



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA
REDE NORDESTE DE BIOTECNOLOGIA – RENORBIO

BRUNO SAMPAIO

**ANÁLISE DAS FREQUÊNCIAS ALÉLICAS E DOS
PARÂMETROS POPULACIONAIS DE INTERESSE FORENSE DE 23
MARCADORES STRs AUTOSSÔMICOS E DE 23 REGIÕES STRs DO
CROMOSSOMO (Y-STRs) NA POPULAÇÃO DO ESTADO DE PERNAMBUCO**

Recife
2022

BRUNO SAMPAIO

**ANÁLISE DAS FREQUÊNCIAS ALÉLICAS E DOS
PARÂMETROS POPULACIONAIS DE INTERESSE FORENSE DE 23
MARCADORES STRs AUTOSSÔMICOS E DE 23 REGIÕES STRs DO
CROMOSSOMO (Y-STRs) NA POPULAÇÃO DO ESTADO DE PERNAMBUCO**

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia (Rede Nordeste de Biotecnologia – RENORBIO) da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de doutor em Biotecnologia. Área de concentração: Saúde.

Orientador: Prof. Dr. Valdir de Queiroz Balbino

Coorientadora: Profa. Dra. Anna Theresa de Souza Liberal

Recife

2022

Catálogo na Fonte:
Bibliotecário Bruno Márcio Gouveia, CRB-4/1788

Sampaio, Bruno

Análise das frequências alélicas e dos parâmetros populacionais de interesse forense de 23 marcadores STRs autossômicos e de 23 regiões STRs do cromossomo (Y-STRs) na população do Estado de Pernambuco / Bruno Sampaio. - 2022.

95 f. : il.

Orientador: Prof. Dr. Valdir de Queiroz Balbino.

Coorientadora: Profa. Dra. Anna Theresa de Souza Liberal

Tese (doutorado) – Universidade Federal de Pernambuco. Centro de Biociências. Programa de Pós-Graduação em Programa de Pós-graduação em Biotecnologia (Rede Nordeste de Biotecnologia – RENORBIO), Recife, 2022.

Inclui referências.

1. Genética das populações. 2. Genética da população humana. 3. Genética forense. I. Balbino, Valdir de Queiroz (orientador). II. Liberal, Anna Theresa de Souza (coorientadora). III. Título.

599.935

CDD (22.ed.)

UFPE/CB-107-2022

BRUNO SAMPAIO

**ANÁLISE DAS FREQUÊNCIAS ALÉLICAS E DOS
PARÂMETROS POPULACIONAIS DE INTERESSE FORENSE DE 23
MARCADORES STRs AUTOSSÔMICOS E DE 23 REGIÕES STRs DO
CROMOSSOMO (Y-STRs) NA POPULAÇÃO DO ESTADO DE PERNAMBUCO**

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia (Rede Nordeste de Biotecnologia – RENORBIO) da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de doutor em Biotecnologia. Área de concentração: Saúde.

Aprovada em 08/03/2022

BANCA EXAMINADORA

Prof. Dr. Valdir de Queiroz Balbino
Universidade Federal de Pernambuco

Profa. Dra. Maria Tereza Cartaxo Muniz
Universidade de Pernambuco

Profa. Dra. Silvia Helena Barem Rabenhorst
Universidade Federal do Ceará

Prof. Dr. Tercilio Calsa Júnior
Universidade Federal de Pernambuco

Prof. Dr. Rafael Lima Guimarães
Universidade Federal de Pernambuco

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Valdir Balbino, pela confiança, paciência e oportunidade de realização deste trabalho.

À Dra. Theresa Liberal, minha coorientadora, por todo o auxílio para o desenvolvimento desta pesquisa.

Aos companheiros do Laboratório de Bioinformática e Biologia Evolutiva (LABBE): a Abigail Marcelino e Bárbara Pereira, em especial, pela amizade, apoio e contribuições para a conclusão deste trabalho; aos integrantes do grupo de Genética Forense (Bruno Almeida, Jayane e Ellen) pela grande ajuda, sem a qual este trabalho não teria sido realizado; a Sérgio Paiva, pelo auxílio nas análises estatísticas; e a Marcos e Heidi, pelos conselhos, apoio e troca de experiências.

À minha mãe, Mara, que sempre me apoiou e foi fundamental para o meu desenvolvimento pessoal e profissional.

À minha esposa, Isla, por todo o amor, companheirismo e incentivo durante essa jornada.

À minha família.

À UFPE e ao Programa de Pós-Graduação em Biotecnologia da Rede Nordeste em Biotecnologia (RENORBIO), pela oportunidade de desenvolver este trabalho e permissão de realização do curso de doutorado.

RESUMO

[illegible]

capacidade discriminação, e probabilidade de coincidência haplotípica foram 0,9998, 0,9542 e 0,00023, respectivamente. A diversidade gênica variou de 0,5380 (DYS393) a 0,9237 (DYS385). Variantes alélicas foram observadas na amostragem, como microvariantes, duplicações em regiões de cópia única e a presença de um alelo nulo (marcador DYS533). Os resultados dos testes de distância genética (F_{ST} e R_{ST}), fundamentados nos Y-STRs, revelaram que a população analisada é geneticamente próxima de outras populações brasileiras e europeias, o que corrobora com a história da colonização do Estado de Pernambuco. Por meio dos resultados obtidos, foi possível concluir que ambos os painéis de 23 marcadores são altamente informativos, sendo assim, bastante úteis para os exames de identificação humana na população estudada.

Palavras-chave: Genética Forense; Identificação Humana; Repetições Curtas em *Tandem*; Genética de Populações; Pernambuco

[illegible]

results obtained, it was possible to conclude that both panels of 23 markers proved to be highly informative, being useful for forensic purposes in this population.

Keywords: Forensic Genetics; Human Identification; Short Tandem Repeats; Population Genetics; Pernambuco State.

LISTA DE ILUSTRAÇÕES

REVISÃO DA LITERATURA

Figura 1 – Representação dos marcadores de tamanho — minissatélites (VNTRs) e microssatélites (STRs).....	24
Figura 2 – Representação da posição relativa dos marcadores STR do CODIS.....	26
Figura 3 – Modelo de diferenciação dos cromossomos sexuais heteromórficos X e Y.....	30
Figura 4 – Representação esquemática das regiões do cromossomo Y humano....	32
Figura 5 – Posições relativas dos marcadores incluídos no kit PowerPlex® Y23 (Promega Corporation).....	35

ARTIGO 1 – GENETIC CHARACTERIZATION OF 23 Y CHROMOSOME SHORT TANDEM REPEATS IN THE STATE OF PERNAMBUCO, BRAZIL

Supplementary figure 1 – (A) Pernambuco State, located in the Brazil's Northeast Region. (B) The five mesoregions of Pernambuco: Recife Metropolitan Region, Zona da Mata, Agreste, Sertao and São Francisco.....	62
Supplementary figure 2 – Multidimensional scale (MDS) plot of the pairwise F_{ST} genetic distances, calculated between Pernambuco and other populations (Stess: 0.01685522).....	63
Supplementary figure 3 – Multidimensional scale (MDS) plot of the pairwise R_{ST} genetic distances, calculated between Pernambuco and other populations.....	64
Supplementary figure 4 – Multidimensional scale (MDS) plot of the pairwise R_{ST} genetic distances, calculated between Brazilian and Central Portugal populations...	65

LISTA DE TABELAS

REVISÃO DA LITERATURA

Tabela 1 – Informações referentes aos marcadores designados para o CODIS.....	28
Tabela 2 – Parâmetros estatísticos forenses relativos aos marcadores STRs autossômicos.....	41
Tabela 3 – Parâmetros estatísticos forenses relativos às Y-STRs.....	42

ARTIGO 1 – GENETIC CHARACTERIZATION OF 23 Y CHROMOSOME SHORT TANDEM REPEATS IN THE STATE OF PERNAMBUCO, BRAZIL

Supplementary table 1 – Comparison about number of haplotypes and forensic parameters considering four set of Y-STRs markers.....	66
Supplementary table 2 – Allele frequencies and gene diversity for the 23 Y-STRs in Pernambuco population.....	67
Supplementary table 3 – Allelic variants observed in the present study.....	69
Supplementary table 4 – F_{ST} genetic distances (under the diagonal) between populations of the five mesoregions of Pernambuco state and non-differentiation correspondent p-values (0.005) (above the diagonal), considering 10,000 permutations.....	69
Supplementary table 5 – List of included populations in the F_{ST} analysis of the present study.....	70
Supplementary table 6 – F_{ST} genetic distances (under the diagonal) between Pernambuco population and other reported populations and non-differentiation correspondent p-values (0.00024, after Bonferroni's correction) (above the diagonal), considering 10,000 permutations.....	71
Supplementary table 7 – List of included populations in the R_{ST} analysis of the present study.....	72
Supplementary table 8 – R_{ST} genetic distances between (under the diagonal) between Pernambuco population and other reported populations and non-differentiation correspondent p-values (0.00017, after Bonferroni's correction) (above the diagonal), considering 10,000 permutations.....	74

ARTIGO 2 – ALLELIC FREQUENCIES DISTRIBUTION AND FORENSIC
PARAMETERS OF 23 AUTOSOMAL SHORT TANDEM REPEATS IN THE
POPULATION OF THE STATE OF PERNAMBUCO, BRAZIL

Table 1 – Statistical data about populational parameters of forensic relevance of 23 autosomal STRs for the Pernambuco population.....	81
--	----

LISTA DE ABREVIATURAS E SIGLAS

AZF	Fator de azoospermia
BNPG	Banco Nacional de Perfis Genéticos
CD	Capacidade de discriminação
CIP	Conteúdo de informação de polimorfismo
CODIS	Sistema Combinado de Indexação de DNA
DG	Diversidade gênica
DH	Diversidade haplotípica
DNA	Ácido desoxirribonucleico
ESS	Conjunto-Padrão Europeu
EHW	Equilíbrio de Hardy-Weinberg
FBI	Departamento Federal de Investigação dos Estados Unidos
H _{Esp.}	Heterozigosidade esperada
HLA	Antígeno Leucocitário Humano
HM	Haplótipo Mínimo
H _{Obs.}	Heterozigosidade observada
IBGE	Instituto Brasileiro de Geografia e Estatística
IP	Índice de paternidade
IP _{típico}	Índice de paternidade típico
ISFG	Sociedade Internacional de Genética Forense
LABBE	Laboratório de Bioinformática e Biologia Evolutiva
mtDNA	DNA mitocondrial
MDS	Escala multidimensional
MSY	Região masculino-específica do cromossomo Y
NDIS	Sistema Nacional de Indexação de DNA
NRY	Região não-recombinante do cromossomo Y
PCR	Reação em Cadeia da Polimerase
PAR	Região pseudoautossômica
PD	Poder de discriminação
PE	Poder de exclusão
PM	Probabilidade de <i>match</i>
PMH	Probabilidade de <i>match</i> haplotípico

RFLP	Polimorfismo no Comprimento de Fragmento de Restrição
SSR	Repetição de Sequência Simples
SRY	Região determinante do sexo do cromossomo Y
SWGDAM	Grupo de Trabalho Científico sobre Métodos do DNA dos Estados Unidos
STR	Repetição Curta em <i>Tandem</i>
UFPE	Universidade Federal de Pernambuco
VNTR	Repetição em <i>Tandem</i> de Número Variável
YHRD	Banco de Dados de Referência de Haplótipos do Cromossomo Y
Y-SNP	Polimorfismo de Nucleotídeo Único do cromossomo Y
Y-STR	Repetição Curta em <i>Tandem</i> do cromossomo Y

SUMÁRIO

1	INTRODUÇÃO.....	16
2	REVISÃO DA LITERATURA.....	18
2.1	GENÉTICA FORENSE.....	18
2.2	IDENTIFICAÇÃO HUMANA PELO DNA.....	21
2.2.1	Organização do Genoma Humano.....	21
2.2.2	Marcadores Genéticos de Identificação Humana.....	21
2.2.2.1	Minissatélites.....	22
2.2.2.2	Microssatélites.....	23
2.3	CROMOSSOMO Y.....	29
2.4	MARCADORES GENÉTICOS DO CROMOSSOMO Y.....	33
2.5	GENÉTICA DE POPULAÇÕES HUMANAS.....	36
2.6	ORIGEM DA POPULAÇÃO BRASILEIRA.....	43
3	OBJETIVOS.....	46
3.1	GERAL.....	46
3.2	ESPECÍFICOS.....	46
4	ARTIGO 1 – GENETIC CHARACTERIZATION OF 23 Y CHROMOSOME SHORT TANDEM REPEATS IN THE STATE OF PERNAMBUCO, BRAZIL.....	48
5	ARTIGO 2 – ALLELIC FREQUENCIES DISTRIBUTION AND FORENSIC PARAMETERS OF 23 AUTOSOMAL SHORT TANDEM REPEATS IN THE POPULATION OF THE STATE OF PERNAMBUCO, BRAZIL.....	75
6	CONCLUSÕES.....	87
	REFERÊNCIAS.....	88
	ANEXO A – COMPROVANTE DE SUBMISSÃO DE ARTIGO À REVISTA AMERICAN JOURNAL OF HUMAN BIOLOGY.....	95
	ANEXO B – COMPROVANTE DE SUBMISSÃO DE ARTIGO À REVISTA LEGAL MEDICINE.....	96

1 INTRODUÇÃO

A utilização de métodos confiáveis na identificação humana sempre foi indispensável à sociedade. Seja para fins sociais, civis, administrativos, comerciais ou penais, o valor da identificação sempre foi indiscutível e imprescindível (FRANÇA, 2011). Em 1985, Alec Jeffreys e colaboradores desenvolveram uma metodologia que possibilitou analisar diretamente o DNA humano, denominada de “impressão digital” do DNA (*DNA fingerprint*) (JEFFREYS; WILSON; THEIN, 1985a,b). Desde então, com o avanço tecnológico, essa técnica sofreu inúmeros melhoramentos e aperfeiçoamentos e atualmente é denominada de *DNA profiling* (ROEWER, 2013).

As aplicações do *DNA profiling* consistem na análise de polimorfismos do DNA para a elucidação de problemas jurídicos oriundos dos âmbitos cível e criminal. Dentre as utilizações do DNA nas análises forense, as principais são as investigações criminais, os testes de paternidade/maternidade e o exame de reconstituição de parentesco biológico (PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014). A utilização do DNA como fonte de evidência forense possibilita um elevado grau de individualização, tendo como fato que 0,5% do genoma é exclusivo para cada indivíduo (PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014).

Os marcadores de identificação humana mais empregados nos laboratórios forenses são os *loci* de STRs autossômicos, considerados marcadores “padrão-ouro” e de “primeira linha”. O *Combined DNA Index System* (CODIS), que combina a tecnologia eletrônica com a biologia molecular, foi implementado pelo *Federal Bureau of Investigation* (FBI) e preconiza o uso de um padrão mínimo de 20 STRs autossômicos, designado como CODIS expandido (FBI, 2015).

As STRs presentes no cromossomo Y, denominadas de Y-STRs, tornaram-se uma ferramenta complementar, no cenário forense, às análises dos STRs autossômicos, sendo utilizadas em diversos casos específicos, especialmente na resolução dos crimes sexuais (PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014). As STRs autossômicas e as Y-STRs já são bastante estudadas, estabelecidas, difundidas e padronizadas mundialmente. Todavia, para que esses marcadores do DNA possam ser utilizados pelos laboratórios forenses criminais e cíveis, é fundamental que estejam validados em determinada população local, pois essas

análises dependem do conhecimento das frequências alélicas e haplotípicas para a realização da estatística forense (BUTLER, 2010; GETTINGS *et al.*, 2018).

2 REVISÃO DA LITERATURA

2.1 GENÉTICA FORENSE

A Genética Forense é um ramo das ciências forenses, definida como sendo a aplicação da Genética e da Biologia Molecular na resolução de conflitos legais, sejam estes de natureza cível ou criminal (CARRACEDO, 2013). Os testes de identificação humana, baseados na variabilidade genética existente nos indivíduos, são as principais ferramentas dessa ciência, sendo empregados principalmente em dois tipos de casuística: i) determinação de vínculo genético entre indivíduos (exames de paternidade, maternidade e de reconstituição de parentesco); e ii) vinculação, ou exclusão, de um ou mais indivíduos a um crime.

A história da Genética Forense foi primeiramente direcionada pelos avanços da hemogenética. Em 1901, Karl Landsteiner, ao descrever os polimorfismos do sistema ABO humano, percebeu que os indivíduos poderiam ser categorizados em grupamentos sanguíneos e que essa variabilidade genética poderia ser empregada para propósitos forenses (SCHWARZ; DORNER, 2003; GESERICK; WIRTH, 2012). Posteriormente, em 1910, Emil von Dungern e Ludwig Hirschfeld lançaram as bases para o uso do sistema ABO nas investigações de paternidade, após o entendimento de que esses marcadores eram herdados conforme as leis mendelianas (GESERICK; WIRTH, 2012).

Após o sistema ABO, novos marcadores genéticos foram descobertos e passaram a ser empregados nas análises forenses: tipagem de grupamentos sanguíneos (sistemas MN, P, Rhesus); polimorfismos de proteínas séricas (haptoglobina, transferrina); sistema de antígeno leucocitário humano (HLA); e polimorfismos de enzimas eritrocitárias (GESERICK; WIRTH, 2012). Embora muitos desses sistemas fossem ferramentas poderosas, as técnicas sorológicas eram limitadas em muitos casos forenses pelo tipo (sangue) e quantidade de material biológico necessária para geração de resultados elevadamente discriminativos (GOODWIN; LINACRE; HADI, 2011).

O advento de novas técnicas de biologia molecular, incluindo o uso de enzimas de restrição, *Southern blotting* e sequenciamento de nucleotídeos, possibilitou a

análise direta de variações contidas das sequências de DNA (SOUTHERN, 1975; SANGER; NICKLEN; COULSON, 1977; GOODWIN; LINACRE; HADI, 2011). A primeira região hipervariável do DNA foi descoberta por meio de fragmentos produzidos pela técnica de Polimorfismo no Comprimento de Fragmento de Restrição [RFLP, do inglês *Restriction Fragment Length Polymorphism*] (WYMAN; WHITE, 1980). Botstein *et al.* (1980) sugeriram a utilização dessa técnica na construção de mapas genômicos que seriam utilizados para a detecção de regiões que pudessem ser empregadas em estudos de variabilidade genética humana.

Em 1985, Alec Jeffreys e colaboradores revolucionaram a Genética Forense com o desenvolvimento de um método de análise direta do DNA humano, baseado na utilização de RFLP para a detecção de regiões minissatélites polimórficas no DNA humano. Nesse método, enzimas de restrição foram utilizadas para clivar o DNA em segmentos menores, gerando os fragmentos de diferentes tamanhos, que foram separados eletroforéticamente em gel de agarose, e depois transferidos por *Southern blotting* para uma membrana de náilon. O DNA fixado na membrana foi hibridizado com uma sonda radioativa, e depois revelado por raios-x, produzindo um padrão de múltiplas bandas. Essa técnica ficou conhecida como “impressão digital” do DNA [DNA *fingerprint*] (JEFFREYS; WILSON; THEIN, 1985a,b; JOBLING; GILL, 2004).

O valor da probabilidade de *match* (correspondência) foi estimado em cerca de 3×10^{-11} , fazendo uso de apenas uma sonda única, e 5×10^{-19} , quando duas sondas eram utilizadas, valores tão ínfimos que apenas gêmeos monozigóticos poderiam compartilhar as mesmas impressões digitais de DNA geradas (JEFFREYS; WILSON; THEIN, 1985b; JOBLING; GILL, 2004). Logo após a descoberta, o método de impressão digital do DNA foi utilizado pela primeira vez para a realização de um teste de determinação de maternidade vinculado a um caso de imigração e, pouco tempo depois, na elucidação de um caso de duplo homicídio (JEFFREYS; BROOKFIELD; SEMEONOFF, 1985; JOBLING; GILL, 2004; JOBLING, 2013).

Concomitantemente a esta descoberta, as ciências forenses foram profundamente beneficiadas com a descoberta da Reação em Cadeia da Polimerase (PCR, do inglês *Polimerase Chain Reaction*), em 1985, por Kary Mullis. Com a capacidade de produzir milhões de cópias de uma sequência específica do DNA, em um espaço relativamente curto de tempo, a PCR possibilitou a análise de DNA a partir

de vários tipos de amostras biológicas, notadamente as de cenas de crime, nas quais o DNA costuma ser limitado qualitativa e quantitativamente (BUTLER, 2010).

Com o melhoramento das técnicas de análise de DNA e a aceitação dessa metodologia pelos tribunais, o Departamento Federal de Investigação dos Estados Unidos (FBI, do inglês *Federal Bureau of Investigation*) criou, em 1988, o CODIS (*Combined DNA Index System* ou Sistema Combinado de Indexação de DNA). Este é um programa nacional que combina ciência forense e informática em uma ferramenta de vinculação de crimes violentos, com o intuito de auxiliar o NDIS (*National DNA Index System* ou Sistema Nacional de Indexação de DNA), um conjunto de bancos de dados de abrangência local, estadual ou nacional, que permite que os laboratórios forenses permutem e comparem perfis de DNA eletronicamente (GODINHO, 2014; FBI, 2021).

No Brasil, a Genética Forense foi norteadada com a inauguração, em 1995, do Laboratório de DNA Criminal da Polícia Civil do Distrito Federal, que ficara responsável pelos treinamentos dos laboratórios de perícia criminal posteriormente implantados em outros estados (Paraíba, Rio Grande do Sul, Minas Gerais, Paraná, Mato Grosso do Sul e São Paulo). Com os incentivos do Programa de Ciência e Tecnologia Aplicada na Segurança Pública, lançado em 2014 em uma ação conjunta do Ministério da Ciência e Tecnologia com a Secretaria Nacional de Segurança Pública, houve uma intensa capacitação de profissionais e a construção de novos laboratórios forenses em diversos estados (SANTOS, 2014).

Em 2010, o CODIS foi cedido à Polícia Federal e inicialmente instalado em Brasília (Distrito Federal). Dois anos após, mediante a criação de uma rede organizada e bem disseminada de laboratórios periciais criminais e à implementação do Banco Nacional de Perfis Genéticos (BNPG), o Brasil começou, em 2012, a alimentar, por meio do CODIS, a sua Rede Integrada de Perfis Genéticos — RIBPG (SILVA *et al.*, 2020).

2.2 IDENTIFICAÇÃO HUMANA PELO DNA

2.2.1 Organização do Genoma Humano

O genoma humano pode ser definido como sendo o complexo da informação genética armazenada no núcleo celular (DNA genômico) e nas mitocôndrias [DNA mitocondrial ou mtDNA] (PIOVESAN *et al.*, 2019). O genoma nuclear somático ($2n$) possui uma extensão de aproximadamente seis bilhões de pares de nucleotídeos, e está organizado em 46 cromossomos (23 pares homólogos), sendo 23 maternos e 23 paternos. A maior parte das células de um indivíduo são diploides e possuem essencialmente o mesmo DNA genômico; as células gaméticas (haploides) e os eritrócitos (não possuem DNA) são as exceções (PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014).

A maior parte do genoma humano (~98%) é representada pelas sequências não-codificantes, que não fazem parte das sequências proteicas e que no passado foram consideradas não-funcionais, sendo rotuladas como “DNA lixo”, mas que atualmente são conhecidas por apresentarem papéis na regulação de genes e no processamento de RNA, além de fornecerem um registro histórico da evolução do genoma (ALEXANDER *et al.*, 2010). Por outro lado, as sequências gênicas codificantes de proteínas (éxons) representam apenas 2% do genoma humano, estando presentes nos cerca de 20.000 genes (LIEHR, 2021).

A variação genética parece ser bastante limitada nas regiões codificantes (exceção feita à região HLA), o que poderia ser explicado por uma maior pressão seletiva a qual os genes expressos estão submetidos durante o processo evolutivo, a fim de garantir que as suas funções sejam preservadas. Em contrapartida, a parte não-codificante aparenta sofrer uma menor pressão seletiva e, consequentemente, as mutações nessa região vão sendo geralmente mantidas e transmitidas, levando ao aumento na variabilidade genética humana (CARRACEDO; SÁNCHEZ-DIZ, 2005).

2.2.2 Marcadores Genéticos de Identificação Humana

As diferenças genéticas entre os seres humanos são mínimas. Do total de pares de base do genoma humano, a análise forense é fundamentada no fato de que

apenas 0,5% do DNA é exclusivo para cada indivíduo (PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014). Dessa forma, a utilização do DNA como fonte de evidência forense possibilita um alto nível de individualização (BUTLER, 2010).

Na espécie humana, os cromossomos homólogos são bastante similares entre si, em sequências e genes, mas variam em regiões específicas (*loci*). Se uma determinada variação for encontrada na população em uma frequência entre 1% e 99%, será denominada de polimorfismo (BEIGUELMAN, 2008). Por conta dessa variação intrapopulacional, os polimorfismos podem ser empregados como eficientes marcadores genéticos de identificação humana (ROCHA *et al.*, 2007).

Uma porcentagem significativa do DNA não-codificante (~30%) consiste em sequências repetitivas, que podem ser divididas em duas classes: sequências repetitivas consecutivas (em *tandem*) e elementos intercalares. Os sistemas de tipagem forense são majoritariamente baseados em *loci* genéticos localizados nas sequências de DNA repetitivas em tandem (CARRACEDO; SÁNCHEZ-DIZ, 2005). Essas sequências possuem tipos de tamanhos variados e são categorizadas conforme o comprimento da unidade (igualmente denominada núcleo ou motivo) de repetição e pela quantidade de vezes que o núcleo se repete continuamente. Estas variações determinam a existência dos diferentes tipos de alelos (BUTLER, 2007).

2.2.2.1 Minissatélites

O primeiro método de impressões digitais do DNA consistia no emprego da técnica de RFLP para a detecção de marcadores minissatélites. Também designados de Repetições em Tandem de Número Variável (VNTRs, do inglês *Variable Number Tandem Repeats*), esses marcadores de tamanho (Figura 1) consistem em sequências de DNA que apresentam unidades de repetição compostas por 10 a 100 pares de base (pb) que se repetem consecutivamente e formam alelos que variam de 500 a 20.000 bp (BUTLER, 2005; CARRACEDO; SÁNCHEZ-RIZ, 2005).

As VNTRs foram intensamente utilizadas na análise forense durante anos, especialmente pela grande variabilidade alélica e, como consequência dessa característica, pelo elevado poder discriminatório. No entanto, devido às limitações decorrentes dos tamanhos dos fragmentos necessários (cerca de 50 ng de DNA é requerido para a análise por RFLP), a sua aplicabilidade na análise de amostras (as

degradadas não podiam ser analisadas) era bastante reduzida (sem contar as possíveis falhas técnicas dos procedimentos de detecção) e, por conta desses fatores, acabaram sendo substituídas como marcadores genéticos forenses (BUDOWLE *et al.*, 1998; CHAKRABORTY *et al.*, 1999; PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014).

2.2.2.2 Microssatélites

As regiões microssatélites, igualmente conhecidas como Repetições Curtas em Tandem (STRs, do inglês *Short Tandem Repeats*) ou Repetições de Sequência Simples (SSRs, do inglês *Simple Sequence Repeats*), são trechos de DNA compostos por motivos de repetição formados por dois a sete pares de bases e que se repetem consecutivamente dezenas de vezes, formando segmentos que variam entre 100 e 500 bp (isso é na análise) (FAN; CHU, 2007; BUTLER, 2007).

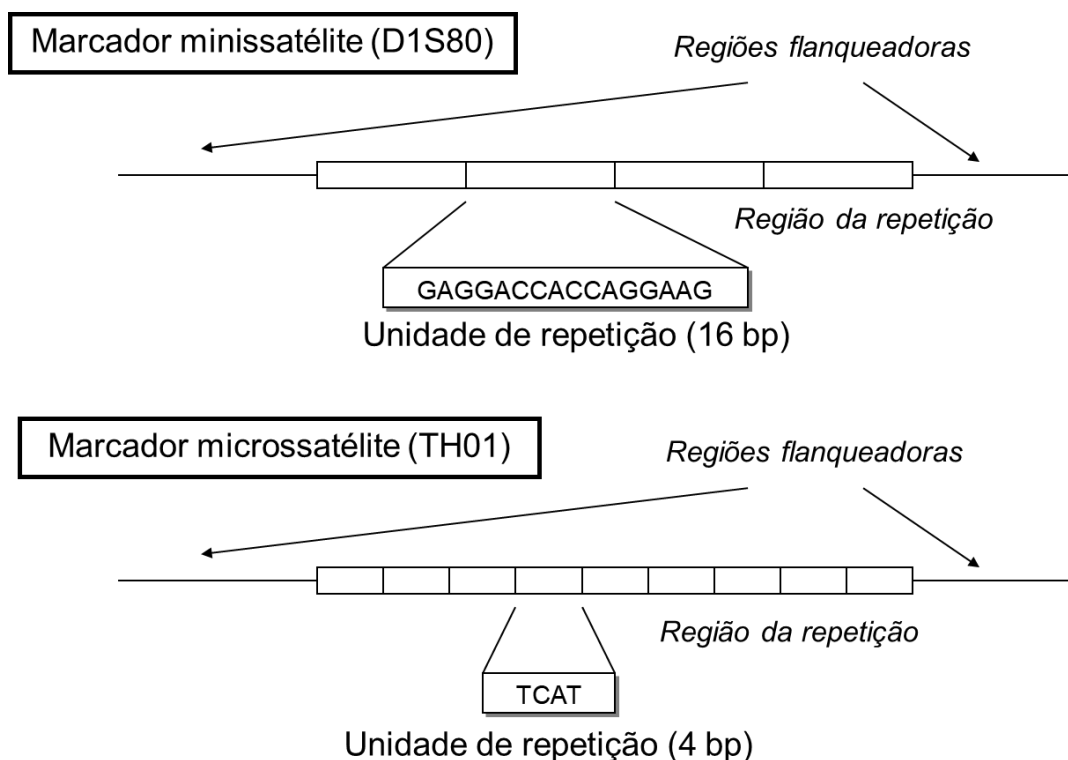
Os microssatélites são amplamente frequentes e distribuídos (representam 3% do genoma humano), apesar de essa distribuição não ser uniforme entre os cromossomos (SUBRAMANIAN *et al.*, 2003; FAN; CHU, 2007). As STRs são mais frequentes em regiões não-codificantes, embora também possam ser encontradas em regiões codificantes (FAN; CHU, 2007). Classificam-se, conforme o número de nucleotídeos que formam a unidade de repetição, em: di (2 bp); tri (3 pb); tetra (4 pb); penta (5 bp); hexa (6 bp); e hepta (7 bp) nucleotídeos (JIN; ZHONG; CHAKRABORTY, 1994; BUTLER, 2005). Os *loci* tetranucleotídicos tornaram-se mais popularizados, devido à menor geração de *stutters* em comparação aos di e trinucleotídicos (BUTLER; 2012). Os *stutters* são produtos de PCR gerados por *slippage* (derrapagem) da DNA polimerase, com uma ou mais unidades de repetição a menos (ou a mais) que o alelo verdadeiro e que podem aumentar, por exemplo, a complexidade da análise de misturas [presença de materiais genéticos oriundos de dois ou mais doadores] (BIEBER *et al.*, 2016).

As STRs podem ainda ser categorizadas, conforme o padrão de repetição, em: simples (contém unidades idênticas em sequência e tamanho); compostas (compreendem dois ou mais tipos de repetições simples adjacentes; e complexas [podem conter vários blocos com unidades de repetição variáveis, associadas a

sequências intervenientes, com alelos que divergem em tamanho e sequência] (BUTLER, 2010).

Com o desenvolvimento da PCR, os sistemas de tipagem de DNA baseados em VNTRs/RFLP acabaram sendo substituídos pelos de STRs/PCR (PARSON, 2018). Os alelos menores das STRs, em comparação aos das VNTRs, aumentaram consideravelmente a sensibilidade da detecção, permitindo a obtenção de perfis genéticos, mesmos em situações com amostras degradadas (JOBLING; GILL, 2004; PARSON, 2018).

Figura 1 – Representação dos marcadores de tamanho — minissatélites (VNTRs) e microsatélites (STRs).



Fonte: Adaptado de Butler (2005).

Os primeiros kits comerciais amplamente usados para a análise de várias STRs autossômicas em uma única reação se tornaram disponíveis no final dos anos 1990 e início dos anos 2000 (VAN OORSCHOT; BALLANTYNE; MITCHEL, 2010).

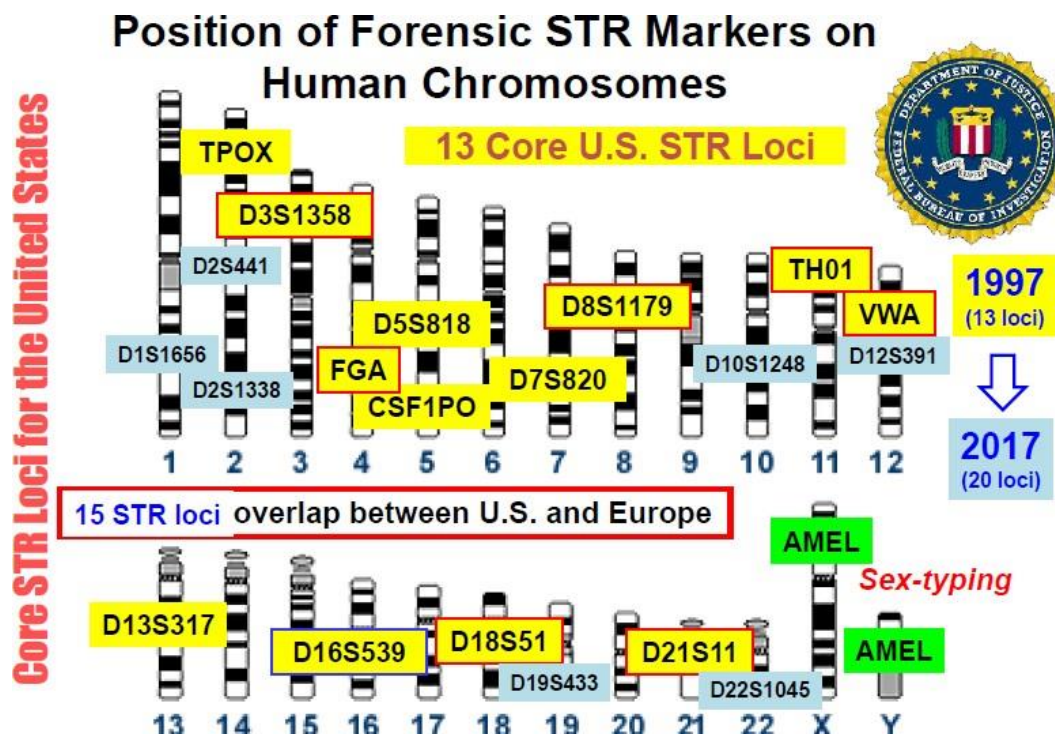
Em 1997, o FBI estabeleceu o primeiro conjunto de marcadores STRs

autossômicos, com o intuito de padronizar a implementação dessas regiões genômicas nas rotinas do CODIS, sendo composto por 13 *loci*: D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, CSF1PO, FGA, TH01, TPOX e vWA (BUDOWLE *et al.*, 1998). Cabe ressaltar que sete desses marcadores já estavam inclusos no Conjunto-Padrão Europeu (ESS, do inglês *European Standart Set* (BUDOWLE *et al.*, 1998; RESENDE, 2016).

Recentemente, o FBI anunciou a expansão desse conjunto, preconizando a inclusão de sete novos marcadores (D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 e D22S1045), que agora compreendem um núcleo de 20 STRs autossômicos denominado CODIS expandido [Figura 2] (FBI, 2015; OLDT; KANTHASWAMMY, 2020). O objetivo dessa expansão foi reduzir o número de correspondências casuais, melhorar a questão da compatibilidade internacional e aumentar o poder de discriminação para casos criminais e de pessoas desaparecidas (HARES, 2012; FBI, 2015). As informações genéticas acerca dos marcadores do CODIS encontram-se detalhadas na Tabela 1.

As STRs autossômicas podem ser consideradas como sendo a principal ferramenta forense de identificação humana, sendo a técnica de tipagem desses marcadores, baseada em PCR multiplex em associação à eletroforese capilar (CE), considerada “padrão-ouro” (PARSON, 2018). Essa metodologia é bastante sensível, sendo utilizada em investigações criminais, representadas especialmente pelos testes de confronto de perfis genéticos, bem como nas determinações de relações biológicas entre indivíduos, das quais os testes de paternidade/maternidade e de reconstituição de parentesco biológico são os mais conhecidos (BUTLER, 2005; PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014).

Figura 2 – Representação da posição relativa dos marcadores STR do CODIS.



Fonte: Butler (2015). Disponível em: https://strbase.nist.gov/pub_pres/Butler-DNA-interpretation-AAFS2015.pdf. (Acessado em: 2021).

Os alelos das STRs autossômicas são codominantes. No indivíduo, existem dois alelos, que são herdados de forma mendeliana, sendo um de origem materna, e o outro de origem paterna. Para um determinado *locus*, o indivíduo pode ser considerado heterozigoto — quando apresenta alelos de tamanhos distintos —, ou homozigoto — ambos os alelos são do mesmo tamanho (embora possam possuir sequências divergentes, informação que pode ser obtida por meio de sequenciamento). Na heterozigose, o indivíduo irá apresentar dois picos, indicando a ocorrência de dois alelos diferentes; na situação de homozigose, o indivíduo apresentará apenas um pico, pois ambos os alelos são do mesmo tamanho e estão, desse modo, sobrepostos (SUBRAMANIAN; MISHRA; SINGH, 2003).

Quando várias regiões STRs de um indivíduo são analisadas simultaneamente, o registro gráfico da eletroforese capilar em um analisador genético automático, denominado tecnicamente de eletroferograma, representa um perfil genético ou genótipo (PRIMORAC; SCHANFIELD; MARJANOVIĆ, 2014).

A empregabilidade dos genótipos baseados em STRs autossômicas é

justificada por diversos fatores, como a hipervariabilidade dessas regiões, resultado de uma relativa elevada taxa de mutação ($\sim 10^{-3}$) e pelo padrão de herança mendeliano (BUTLER, 2005). A análise realizada com múltiplos marcadores simultaneamente confere um elevado poder de individualização (GOODWIN; LINACRE; HADI, 2007).

Além dos exames forenses, esses marcadores também podem ser utilizados em estudos populacionais, evolutivos e de ancestralidade (CHAKRABORTY *et al.*, 1999; CALLEGARI-JACQUES *et al.*, 2003; PARSON, 2018).

Tabela 1 – Informações referentes aos marcadores designados para o CODIS.

Locus	Localização cromossômica	Unidade de repetição	Faixa alélica
Amelogenina	X (212 bp), Y (218 bp)	-	-
D3S1358	3p21.31	[AGAT], [TCTA]	8–20
D5S818	5q23.2	[AGAT]	6–18
D7S820	7q21.11	[GATA]	5–16
D8S1179	8q24.13	[TATC]	7–20
D13S317	13q31.1	[GATA]	5–17
D16S539	16q24.1	[GATA]	4–16
D18S51	18q21.33	[GAAA]	7–39.2
D21S11	21q21.1	[TCTA], [TCTG]	12–41.2
CSF1PO	5q33.1; proto-oncogene c-fms humano para o gene do receptor CSF-1, 6º intron	[AGAT]	5–16
FGA	4q28; localizado no 3º intron do gene alfa do fibrinogênio humano	[TTTC]3TTTTTCT[CTTT]nCTCC[TTCC]2	12.2–51.2
TH01	11p15.5; 1º intron do gene da tirosina hidroxilase humana	[TCAT]	3–14
TPOX	2p25.3; 10º intron do gene da peroxidase tireoidiana humana	[AATG]	4–16
vWA	12p13.31; Fator de von Willebrand, 40º intron	[TCTA] [TCTG]	10–25
D1S1656	1q42	[TAGA]n[TGA]0-1[TAGA]n[TAGG]0-1[TG]5	9–21
D2S441	2p14	[TCTA]	8–17
D2S1338	2q35	[TGCC]n[TTCC]n	11–28
D10S1248	10q26.3	[GGAA]n	8–19
D12S391	Cromossomo 12	[AGAT]8-17[AGAC]6-10[AGAT]0-1	15–26
D19S433	19q12	(AAGG)(AAAG)(AAGG)(TAGG)[AAGG]n	5.2–20
D22S1045	22q12.3	[ATT]n	8–20

Fonte: <https://strbase.nist.gov/coreSTRs.htm>

2.3 CROMOSSOMO Y

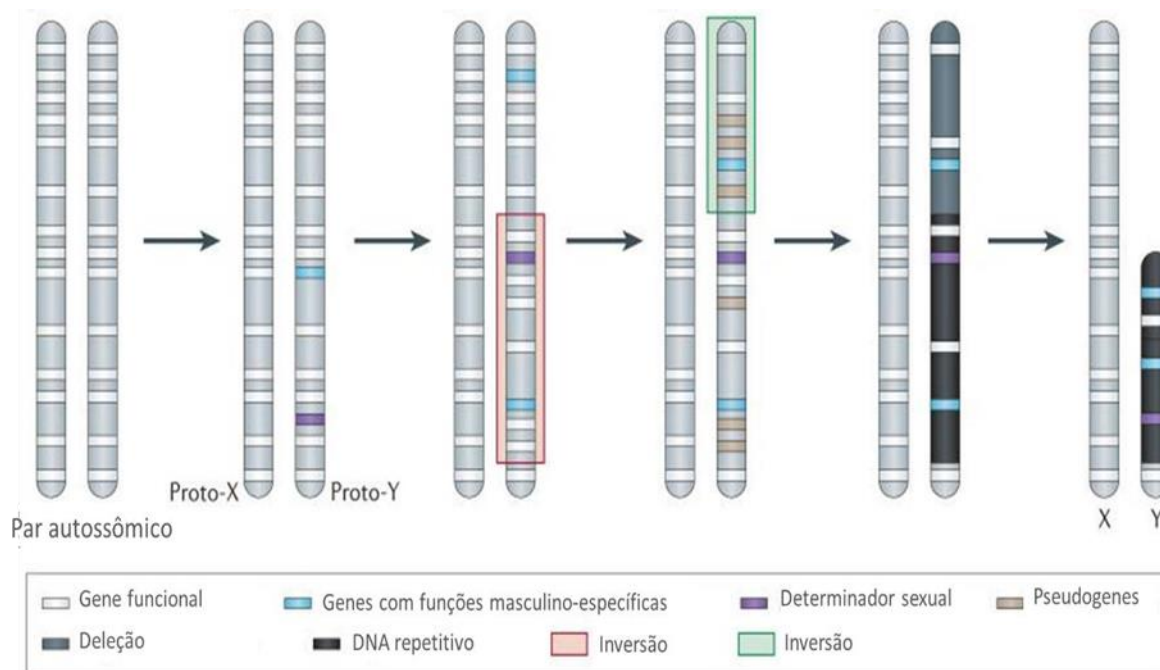
O genoma humano é constituído por 22 pares de cromossomos autossômicos e um par sexual (X e Y). O sexo genético do embrião é determinado no momento da fertilização, com a herança dos cromossomos sexuais presentes nos gametas masculino e feminino. O cromossomo Y hospeda o gene de determinação sexual (SRY, do inglês *sex-determining region Y*), que é responsável por iniciar os processos de formação testicular; a ausência ou disfunção desse gene leva à via do desenvolvimento ovariano (LARNEY; BAILEY; KOOPMAN, 2014).

Os cromossomos sexuais humanos evoluíram a partir de um par ancestral homólogo de cromossomos autossômicos, há mais de 160 milhões de anos, antes, portanto, da divergência de marsupiais e mamíferos euterianos (TROMBETTA; CRUCIANI, 2017). O processo da diferenciação foi iniciado quando houve a aquisição de um gene de determinação sexual por um membro do par autossômico, seguido por processos múltiplos de inversão mediados pela supressão da recombinação, permitindo a evolução independente dos cromossomos protossexuais X e Y [Figura 3] (BATCHROG, 2013; TROMBETTA; CRUCIANI, 2017). Além do acúmulo de outros genes masculinos, a diminuição dos processos recombinantes resultou na perda de função de genes ancestrais (pseudogenização), seguida por degeneração genética e exclusão de grandes segmentos de DNA não essenciais, reduzindo o tamanho físico cromossômico (WATERS; WALLIS; GRAVES, 2007; BATCHROG, 2013).

O resultado desse processo evolutivo levou ao surgimento de cromossomos sexuais heteromórficos, sendo o cromossomo X relativamente grande (155 Mb), assemelhando-se ao seu ancestral autossômico e possuindo cerca de 1.100 genes, envolvidos em funções diversas, além das sexuais (WATERS; WALLIS; GRAVES, 2007).

Em contrapartida, o cromossomo Y consiste em uma pequena estrutura de aproximadamente 60 Mb (~2% do genoma haploide humano), que apresenta a menor quantidade de genes e a maior densidade de sequências repetitivas do genoma nuclear humano (QUINTANA-MURCI; FELLOUS, 2001; WATERS; WALLIS; GRAVES, 2007). É geneticamente dividido em dois conteúdos: um pseudoautossomal e outro exclusivo do sexo masculino (SKALETISKY *et al.*, 2003).

Figura 3 – Modelo de diferenciação dos cromossomos sexuais heteromórficos X e Y.



Fonte: Adaptado de Bachtrog (2013).

As regiões pseudoautossômicas (PAR, do inglês *pseudoautosomal regions*) constituem pequenos trechos localizados nas porções teloméricas do cromossomo Y. A PAR1 está localizada no braço curto (Yp) do cromossomo Y e possui 2,6 Mb de comprimento; a PAR2, com apenas 0,3 Mb, encontra-se localizada na extremidade do braço longo (Yq). As PAR são os únicos domínios do Y que realizam recombinação meiótica (*crossing-over*) com regiões homólogas situadas nas extremidades do cromossomo X, cujo pareamento garante o sucesso da meiose masculina (THOMAS; HASSOLD, 2003; KAUPPI *et al.*, 2011). Os 27 genes presentes nessas regiões estão envolvidos na manutenção celular, são expressos ubiquamente, e possuem correspondentes homólogos no cromossomo X (QUINTANA-MURCI; FELLOUS, 2001; SKALETISKY *et al.*, 2003; NAVARRO-COSTA, 2012).

A maior parte do conteúdo do cromossomo Y (~95%) é geneticamente isolada da recombinação meiótica X-Y, sendo denominada de região não- recombinante do Y (NRY, do inglês *non-recombining region*) (NAVARRO- COSTA, 2012). Skaletsky *et al.* (2003) propuseram que a denominação mais adequada seria região masculino-

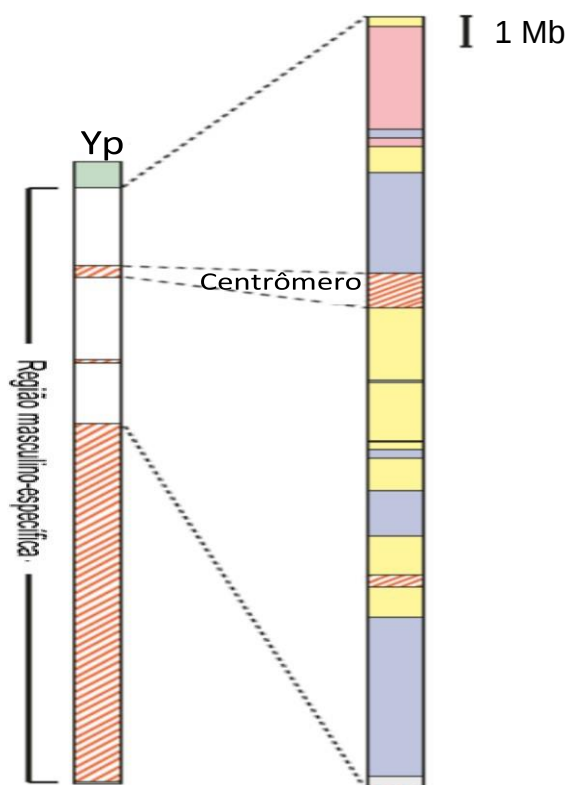
específica (MSY, do inglês *male-specific region*), por conta da evidência da frequente ocorrência de intraconversão gênica ou recombinação intracromossômica.

A MSY é composta pelas regiões eucromáticas e heterocromáticas (Figura 4). Embora essas sequências sejam usualmente tratadas como regiões isoladas, porções de sequências heterocromáticas foram identificadas interrompendo as porções eucromáticas (SKALETISKY *et al.*, 2003). A região eucromática, como a própria designação revela, possui atividade de codificação de proteínas, e está compreendida nos braços curto e longo do cromossomo Y, apresentando 23 Mb de extensão. O sequenciamento desse segmento revelou a existência de 156 unidades de transcrição, das quais somente 78 codificam proteínas, reunidas em 27 famílias gênicas (SKALETISKY *et al.*, 2003).

A região eucromática foi classificada em três tipos de conteúdo: i) as sequências X-transpostas, 99% idênticas às sequências de uma banda (Xq21) localizada no braço longo do cromossomo e são resultado de uma transposição massiva do X para Y, ocorrida há 3-4 milhões de anos. Apenas dois genes foram identificados nessa região de 3,4 Mb de comprimento; ii) X-degeneradas, blocos de segmentos que juntos possuem um comprimento total de 8,4 Mb. Possuem genes e pseudogenes que apresentam entre 60-99% de similaridade de sequências com os respectivos homólogos no cromossomo X, parecendo ser relíquias dos antigos cromossomos autossômicos de origem. O gene SRY está localizado nessa região; iii) amplicônicas, que consistem em sete grandes blocos (10,2 Mb); cerca de 60% de suas sequências exibem similaridade de mais de 99,9% com outras sequências de MSY. Nessa região estão localizadas várias famílias de genes expressos exclusivamente nos testículos (SKALETISKY *et al.*, 2003).

As funções determinantes, do ponto de vista evolutivo, são promovidas pelo gene de determinação sexual (SRY) e pela família de genes fator de azoospermia (AZF, do inglês *azoospermia factor*) (QUINTANA-MURCI; FELLOUS, 2001; SKALETISKY *et al.*, 2003). Embora a determinação sexual e a espermatogênese possuam os papéis mais relevantes do cromossomo Y, genes com outras funções também estão presentes na região eucromática (NAKAHORI; TAKENAKA; NAKAGOME, 1991; SKALETISKY *et al.*, 2003; KOPSIDA *et al.*, 2009).

Figura 4 – Representação esquemática das regiões do cromossomo Y humano.



Fonte: Adaptado de Skaletsky *et al.* (2003).

A região heterocromática é usualmente descrita como uma imensa região contínua de 40 Mb, localizada exclusivamente no Yq. É completamente não-codificante e ainda não foi completamente sequenciada, devido às limitações atuais das técnicas de sequenciamento (CHARLESWORTH, 2003; ALVAREZ- CUBERO *et al.*, 2018). Devido à ausência de recombinação, a MSY é exclusivamente paterna (uniparental) e transmitida clonalmente ao longo das gerações. As sequências presentes nessa região, portanto, só são alteradas pelos eventos mutacionais (QUINTANA-MURCI; FELLOUS, 2001; NAVARRO-COSTA, 2012). Tais características tornam o cromossomo Y extremamente valioso para aplicações forenses. A ausência do Y, no sexo feminino, permite que esse cromossomo possa ser utilizado como um eficiente marcador de identificação sexual. O perfil de herança genética da MSY possibilita o desenvolvimento de ensaios destinados à identificação específica do DNA masculino, o que é imensamente importante nos casos de misturas, especialmente nos crimes sexuais, eliminando a necessidade de realização

de extração diferencial para a separação das frações masculina e feminina (KAYSER; BALLANTYNE, 2012).

2.4 MARCADORES GENÉTICOS DO CROMOSSOMO Y

O cromossomo Y, da mesma maneira que os demais cromossomos humanos, apresenta variações no conteúdo de suas sequências. Dois sistemas principais de marcadores genéticos da MSY podem ser empregados para o estudo da diversidade desse cromossomo: as Y-STRs (multialélicas) e os Y-SNPs (do inglês, *Y-Chromosomal Single Nucleotide Polymorphisms*, ou polimorfismo de nucleotídeo único do cromossomo Y), que são bialélicos (ou binários). Ambos são considerados marcadores de linhagem paterna (ou patrilinearidade); as linhagens maternas (matrilinearidade), por sua vez, são analisadas pelo estudo da sequência do mtDNA (BUTLER, 2010).

Os Y-SNPs são polimorfismos considerados “estáveis”, pois possuem uma baixa taxa de mutação por base por geração [$\sim 10^{-8}$] (BALANOVSKY, 2017). Por isso, são de baixa resolução, sendo mais voltados para estudos de evolução humana (WEI *et al.*, 2013). Por outro lado, as Y-STRs são consideradas “instáveis”, devido à elevada taxa de mutação por base por geração ($\sim 10^{-3}$, similar à das STRs autossômicas). Como consequência, são mais polimórficas e consideradas ideais para as aplicações forenses relacionadas ao cromossomo Y (KAYSER; BALLANTYNE, 2012).

Da maneira similar às STRs autossômicas, as Y-STRs são passíveis de combinação por reação em PCR multiplex e detecção alélica de múltiplos *loci* por eletroforese capilar. Os perfis genéticos de Y-STRs são denominados de haplótipos, por serem transmitidos em bloco (BUTLER, 2005). Devido à essa peculiaridade, um haplótipo de Y-STRs não caracteriza isoladamente um indivíduo, mas sim todos os sujeitos masculinos de uma mesma linhagem paterna (KAYSER; BALLANTYNE, 2012).

As Y-STRs possuem diversas aplicações específicas: investigação de parentesco biológico (quando o pai é ausente ou falecido); resolução de crimes, especialmente os sexuais; identificação de vítimas de desastres; e em estudos históricos, genealógicos e de evolução (ROLF *et al.*, 2001; KAYSER, 2017; CALACAL *et al.*, 2005; FOSTER *et al.*, 1998; ZERJAL *et al.*, 2003; WEI *et al.*, 2013).

A primeira tipagem de uma Y-STR serviu para a exclusão de um suspeito de cometimento de um crime, quando o marcador DYS19 foi utilizado em conjunto a um marcador autossômico (ROEWER; EPPLEN, 1992; ROEWER, 2019). Subsequentemente, diversas Y-STRs foram sendo descobertas e incorporadas nas rotinas dos laboratórios forenses (BALLANTYNE; KAYSER, 2012).

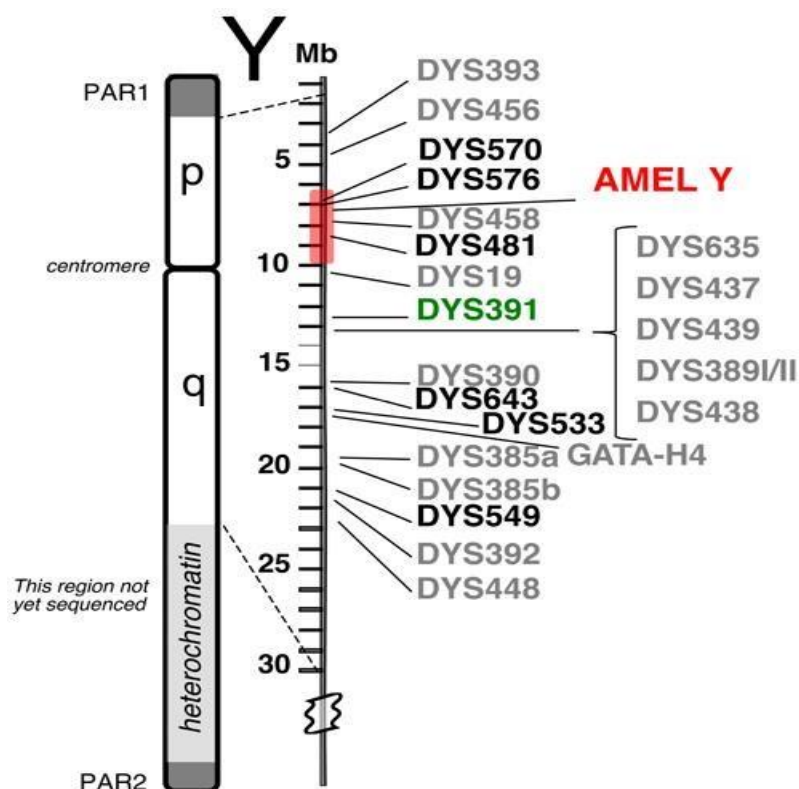
Com o aumento do interesse da comunidade forense na utilização das Y-STRs, um estudo recomendou a aplicação de um conjunto de oito *loci* — que ficou conhecido como “haplótipo mínimo” (HM) — composto por sete marcadores de cópia única (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392 e DYS393) e um de cópia dupla [DYS385a/b] e altamente polimórfico (de KNIJFF, 1997; BUTLER, 2005). Em 2003, o Grupo de Trabalho Científico sobre Métodos do DNA dos Estados Unidos (SWGDM, do inglês *Scientific Working Group on DNA Analysis Methods*) recomendou a adição de dois *loci* de cópia única — DYS438 e DYS439 para as análises forenses e os estudos direcionados ao desenvolvimento de bancos de dados populacionais (SHEWALE, *et al.*, 2004; BUTLER, 2005; THOMPSON *et al.*, 2013).

Diversos kits comerciais têm sido desenvolvidos e empregados com sucesso nas análises forenses envolvendo o cromossomo Y (KAYSER, 2017). Todos são sistemas multiplex bastante sensíveis, capazes de amplificar pequenas frações de DNA masculino, mesmo em amostras ricas em DNA feminino, e que foram padronizados para serem compatíveis com os bancos de dados populacionais (MULERO *et al.*, 2006; THOMPSON *et al.*, 2013). Uma dessas ferramentas comerciais é o kit PowerPlex® Y23 System, desenvolvido e comercializado pela empresa norte-americana Promega Corporation, sendo um sistema multiplex fluorescente de 5 cores que permite a coamplificação 23 Y-STRs (DYS19, DYS385a/b, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, Y-GATA-H4, DYS481, DYS533, DYS549, DYS570, DYS576, e DYS643 [Figura 5] (THOMPSON *et al.*, 2013) e que foi desenhado para ser compatível com os bancos de dados de marcadores do cromossomo Y.

O YHRD (acrônimo de *Y-Chromosome STR Haplotype Reference Database* ou Banco de Dados de Referência de Haplótipos do Cromossomo Y) é o maior banco mundial de dados populacionais de marcadores genéticos do cromossomo Y (acessível pelo endereço eletrônico <https://yhrd.org/>). Esse banco conta com colaborações de 141 países (até o momento), estando em constante expansão e

contendo mais de 340.786 HM e 98.565 haplótipos gerados pela tecnologia PowerPlex® Y23 (ROEWER, 2003; YHRD, 2022).

Figura 5 – Posições relativas dos marcadores incluídos no kit PowerPlex® Y23 (Promega Corporation).



Fonte: <https://www.promega.com.br/products/pm/genetic-identity/population-statistics/power-of-discrimination/>

Embora as Y-STRs sejam analiticamente similares às STRs autossômicas, a interpretação de seus haplótipos é distinta daquela dos genótipos. Quando ocorre, por exemplo, uma correspondência entre os perfis haplotípicos de uma amostra de evidência de cena de crime e uma amostra de um suspeito, não é possível individualizar o suspeito como sendo o doador do material biológico (BALLANTYNE; KAYSER, 2012; KAYSER, 2017).

Essa correspondência deve abranger todos os parentes da mesma patrilinearidade, pois esses também compartilham o mesmo perfil haplotípico (BALLANTYNE; KAYSER, 2012; KAYSER, 2017). Nesse sentido, como a regra do

produto não pode ser aplicada devido ao padrão de herança da MSY, abordagens estatísticas (e.g. método de contagem) irão estimar a raridade do haplótipo dentro de uma determinada população de referência (ROEWER, 2019).

2.5 GENÉTICA DE POPULAÇÕES HUMANAS

A Genética de Populações pode ser definida como o estudo das leis de Mendel e de outros princípios genéticos no contexto da distribuição e variação das frequências alélicas, gênicas e de caracteres hereditários, bem como dos fatores que mantêm ou alteram essas frequências nas populações (BEIGUELMAN, 2008; HARTL; CLARK, 2010).

A Genética de Populações é fundamental para conferir significância estatística aos resultados das análises forenses: os dados de frequências alélicas dos marcadores STRs autossômicos são utilizados diretamente para calcular o peso estatístico de uma evidência genética (GETTINGS *et al.*, 2018). No exame forense mais tradicional, que é o teste de paternidade do tipo trio, os perfis genéticos autossômicos da mãe, filho(a) alegado(a) e suposto pai são comparados para averiguar se o suposto pai é, de fato, o verdadeiro pai biológico.

No caso de uma inclusão, o índice de paternidade (IP) é calculado para cada *loci* analisado. O IP, que costuma ser apresentado pela fórmula X/Y , é definido como sendo a razão de duas probabilidades condicionais: o numerador (X) representa a probabilidade de o suposto pai ter doado o alelo paterno obrigatório, e o denominador (Y) indica a probabilidade de um outro homem qualquer da população ter transmitido esse mesmo alelo (BUTLER, 2005; BARBARO, 2018).

Comumente, X pode ser representado por 1 (100%) [o suposto pai é homozigoto para o alelo de interesse] ou 0,5 (50%) [o suposto pai é heterozigoto], enquanto Y é representado pela frequência alélica do alelo na população. Como o valor de IP é calculado para todos os *loci* de uma determinada análise, os valores de Y dependem da disponibilidade de um banco de dados de frequências alélicas com diversos marcadores (BUTLER, 2005; BARBARO, 2018).

Como essas frequências variam entre populações, preconiza-se o desenvolvimento de bancos apropriados, de preferência representativos para cada população, construídos a partir de estudos genético-populacionais com amostras

preferencialmente randomizadas e representativas, com a finalidade de que os resultados das análises forenses sejam mais acurados (GETTINGS *et al.*, 2018). Vale ressaltar que, além do IP, as demais abordagens estatísticas da Genética Forense são dependentes dos dados de frequências alélicas (BALDING; NICHOLS, 1994; BUCKLETON *et al.*, 2016; ROEWER *et al.*, 2020).

Um dos princípios fundamentais em genética de populações foi estabelecido quando Wilhelm Weinberg e Godfrey Harold Hardy chegaram às mesmas conclusões de maneira independente, e quase que simultaneamente, em 1908. Esse princípio, conhecido como Lei de Hardy-Weinberg, estabelece que quando inexitem fatores evolutivos atuando sobre uma população, as frequências gênicas permanecem constantes ao longo do tempo (BEIGUELMAN, 2008).

Sendo assim, quando uma população obedece a essa lei, é dito que ela está em Equilíbrio de Hardy-Weinberg (EHW), conseqüentemente é possível, a partir das frequências alélicas, determinar as frequências genotípicas (GOODWIN; LINACRE; HADI, 2011).

Para que uma população esteja em EHW, certos critérios devem ser obedecidos (BEIGUELMAN, 2008; GOODWIN; LINACRE; HADI, 2011):

- A população é infinitamente grande;
- Os casamentos ocorrem de maneira randômica (a população é panmítica);
- Não há imigração
- Não há seleção natural;
- E não deve ocorrer mutação.

Como na realidade nenhuma população humana atende simultaneamente a todas essas condições, é necessário investigar se as frequências alélicas e genotípicas se mantêm constantes ao longo das gerações (BEIGUELMAN, 2008; GOODWIN; LINACRE; HADI, 2011). O EHW é determinado para as STRs autossômicas, pois estas obedecem aos princípios genéticos mendelianos, podendo ser calculado por diversas abordagens estatísticas, tais como o teste exato de Fisher e o teste do qui-quadrado (GOODWIN; LINACRE; HADI, 2011; PAIVA JÚNIOR, 2016).

Além do EHW, para a validação do banco de frequências, é necessário estimar a eficiência do conjunto de marcadores para as análises forenses na população em

questão (TILLMAR, 2011). Para as STRs autossômicas, os seguintes parâmetros forenses (Tabela 2) são estimados:

- **Heterozigosidade:** mede a proporção de indivíduos heterozigotos na população, com base em um marcador específico. A heterozigosidade observada ($H_{Obs.}$) de um marcador é obtida de maneira direta, dividindo-se o número de indivíduos heterozigotos pelo número de indivíduos estudados (BUCKLETON *et al.*, 2016). Em contrapartida, a heterozigosidade esperada ($H_{Esp.}$), conhecida alternativamente como diversidade gênica (DG), é estimada com base nas frequências alélicas encontradas na população analisada, por meio da equação de Nei (NEI, 1987 apud BUCKLETON *et al.*, 2016). Uma elevada heterozigosidade revela uma maior diversidade alélica (maior variabilidade genética) e, por conseguinte, uma menor chance de correspondência (*matching*) de perfis. Populações isoladas tendem à homozigosidade devido à endogamia. (BUTLER, 2005). Na comparação entre $H_{Obs.}$ e $H_{Esp.}$, os desvios são indicadores da dinâmica populacional (McMANNUS *et al.*, 2011);
- **Probabilidade de matching (PM):** Significa a probabilidade de dois indivíduos selecionadas ao acaso apresentarem o mesmo genótipo em um determinado *locus* (BUTLER, 2005). Quanto maior a variabilidade genética de um marcador, menor será o valor da sua probabilidade de *matching* (RISCH; DEVLIN, 1992);
- **Poder (ou probabilidade) de discriminação (PD):** Esse parâmetro é o inverso da PM (BUTLER, 2005). É a probabilidade de dois indivíduos selecionados ao acaso apresentarem genótipos distintos em um determinado *locus*. Assim, quanto maior for o PD, melhor será a utilidade forense analítica de um marcador (MOHAMED, 2001);
- **Conteúdo de informação de polimorfismo (CIP):** Representa a probabilidade de os alelos da prole permitirem a dedução dos alelos paternos e maternos. Conforme o valor de CIP, os *loci* podem ser altamente informativos ($CIP > 0,5$), razoavelmente informativos ($0,5 \geq CIP \geq 0,25$) e pouco informativos ($CIP < 0,25$) (BOTSTEIN *et al.*, 1980);
- **Poder de exclusão (PE):** Evidencia a capacidade de um marcador genético em excluir um indivíduo não relacionado, e selecionado randomicamente em

uma determinada população, como suposto pai em uma investigação de paternidade. Por simples multiplicação dos valores de PE de cada *loci*, obtém-se o PE Combinado (PEC), cujo conhecimento pode definir os *loci* a serem utilizados em uma análise genética (OLIVEIRA *et al.*, 2006);

- **Índice de paternidade típico ($IP_{típico}$):** O índice de paternidade (IP) indica quantas vezes é mais provável um suposto pai ser o pai biológico, ao invés de um outro homem selecionado aleatoriamente na população (MOHAMED, 2001). O índice de paternidade típico ($IP_{típico}$) é aquele que é calculado a nível populacional para cada marcador (HUSTON, 1998).

Os parâmetros forenses também são calculados para os marcadores Y-STRs (Tabela 3). Como as Y-STRs são herdadas em bloco, por consequência, um perfil haplotípico é tratado como sendo um alelo único. Sendo assim, todos os parâmetros abaixo avaliam a eficiência forense de um conjunto específico de marcadores, exceto a diversidade gênica, que é estimada a nível de *locus* (PURPS *et al.*, 2014). Os parâmetros estimados nos estudos genético-populacionais são os seguintes:

- **Diversidade haplotípica (DH):** Representa a probabilidade de dois homens não relacionados e selecionados ao acaso na população não compartilharem o mesmo haplótipo. Esse parâmetro revela a variabilidade genética do cromossomo Y existente em uma população (NEI; TAJIMA, 1978). Todavia, o nível de resolução desse parâmetro depende do conjunto de marcadores utilizados (HANSON; BALLANTYNE, 2007; PURPS *et al.*, 2014);
- **Diversidade gênica (DG):** A diversidade gênica é similar à heterozigosidade esperada ($H_{Esp.}$) dos marcadores autossômicos (NEI; TAJIMA, 1981; HANSON; BALLANTYNE, 2007). Tal qual, revela a variabilidade genética de um marcador: quanto maior é a DG de um *loci*, menor será a probabilidade de match randômico entre dois perfis alélicos (NEI; TAJIMA, 1981). A utilidade forense de um *locus* Y-STR na distinção de indivíduos não é apenas determinada por seu valor de DG, mas também está relacionada a quão bem ele complementa os outros marcadores do conjunto de tipagem utilizado (HANSON; BALLANTYNE, 2007);

- **Capacidade de discriminação (CD):** A capacidade de discriminação é definida como a capacidade que um sistema multiplex de Y-STRs possui em discriminar indivíduos do sexo masculino dentro de uma determinada população (HANSON; BALLANTYNE, 2007). A CD representa a razão entre o número de haplótipos diferentes encontrados na amostragem e o número total de haplótipos (indivíduos) analisados (HANSON; BALLANTYNE, 2007; PURPS *et al.*, 2014);
- **Probabilidade de coincidência haplotípico (PMH):** Determina a probabilidade de dois homens randomicamente selecionados em uma população apresentarem o mesmo haplótipo. Trata-se de um parâmetro que é o inverso da DH (BARRA *et al.*, 2014).

Tabela 2 – Parâmetros estatísticos forenses relativos aos marcadores STRs autossômicos.

Parâmetro estatístico	Equação	Descrição
Heterozigosidade esperada ($H_{Esp.}$)	$H_e = 1 - \sum p_i^2$	Onde p_i é a frequência estimada do i -ésimo alelo (McMANNUS <i>et al.</i> , 2011; BUTLER, 2005).
Probabilidade de <i>matching</i> (PM)	$PM_l = \sum_i \tilde{p}_{il}^2$	Onde p_i é a frequência do i -ésimo genótipo do locus l (BUCKLETON <i>et al.</i> , 2016).
Poder de discriminação (PD)	$PD = 1 - \sum_i \tilde{p}_{il}^2$	É o inverso da probabilidade de <i>matching</i> .
Conteúdo de informação de polimorfismo (CIP)	$1 - \left(\sum_{i=1}^n p_i^2 \right) - \sum_{i=1}^{n-1} \sum_{j=i+1}^n 2p_i^2 p_j^2$	Onde n é o número de alelos e p_i e p_j as frequências alélicas (BOTSTEIN <i>et al.</i> , 1980).
Poder de exclusão (PE)	$PE = H_o^2 (1 - 2H_o H^2)$	Onde H_o é a heterozigosidade observada e H é a frequência dos homozigotos (BUTLER, 2005).
Índice de paternidade típico ($IP_{típico}$)	$IP_{típico} = \frac{1}{2H}$	Onde H representa a frequência de homozigotos (HUSTON, 1998)

Fonte: Este trabalho.

Tabela 3 – Parâmetros estatísticos forenses relativos às Y-STRs.

Parâmetro estatístico	Equação	Descrição
Diversidade haplotípica (DH)	$DH = \frac{n}{n-1} (1 - \sum p_i^2)$	Onde n é o número de indivíduos analisados e p_i representa a frequência do i-ésimo haplótipo (NEI, 1987 <i>apud</i> PURPS <i>et al.</i> , 2014).
Diversidade gênica (DG)	$DG = \frac{n}{n-1} (1 - \sum p_i^2)$	Onde n é o número de indivíduos analisados e p_i representa a frequência do i-ésimo alelo (NEI, 1987 <i>apud</i> PURPS <i>et al.</i> , 2014).
Capacidade de discriminação (CD)	$CD = \frac{n}{n_{HDif}}$	Onde n representa o número de indivíduos analisados sobre número de haplótipos diferentes (HANSON; BALLANTYNE, 2007).
Probabilidade de <i>matching</i> haplotípico (PMH)	$PMH = \sum p_i^2$	Onde p_i representa a frequência do i-ésimo haplótipo (PURPS <i>et al.</i> , 2014).

Fonte: Este trabalho.

2.6 ORIGEM DA POPULAÇÃO BRASILEIRA

O Brasil, país de dimensões continentais, foi o berço para a formação de uma das populações mais heterogêneas do mundo, resultante de cinco séculos de cruzamentos interétnicos realizados, principalmente, entres três povos — ameríndios, europeus e africanos (PENA *et al.*, 2011). Os ameríndios foram os primeiros colonizadores do território brasileiro desde o Pleistoceno e, na época do início da colonização portuguesa, compunham uma população de 2,4 milhões de habitantes (CARVALHO-SILVA *et al.*, 2001; IBGE, 2007).

Com o descobrimento das novas terras pelos portugueses pelo Governo de Portugal), os europeus, quase que exclusivamente portugueses, vieram colonizar o novo país (1500–1800). Essa colonização (~500.000 pessoas) envolveu predominantemente homens (CARVALHO-SILVA *et al.*, 2001; IBGE, 2007). De 1630 a 1654, parte do Nordeste brasileiro esteve sob domínio holandês, entretanto, o número de imigrantes de origem holandesa é desconhecido (CARVALHO-SILVA *et al.*, 2001; IBGE 2007).

Durante o período do comércio escravagista, que perdurou oficialmente de 1538 até 1850, estima-se que 3,6 milhões de africanos foram forçados a migrar para o Brasil (PALHA *et al.*, 2012). Após 1808, o Brasil recebera outros seis milhões de imigrantes, a maioria portugueses e italianos, seguidos de espanhóis, alemães, sírios, libaneses e japoneses (CARVALHO-SILVA *et al.*, 2001). Apesar disso, as cinco regiões geopolíticas do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste e Sul) apresentam histórias distintas de colonização e povoamento, fato que se reflete na heterogeneidade genética da população brasileira atual (PALHA *et al.*, 2012; RESQUE *et al.*, 2016).

A Região Nordeste (NE) é uma das cinco macrorregiões que compõe o território brasileiro. É constituída por nove das 27 unidades federativas do Brasil. A população nordestina representa mais de um quinto da população brasileira (IBGE, 2010). O Nordeste, assim como as demais regiões brasileiras, é visto como uma região que apresenta uma população com elevada heterogenia, resultado de um nível acentuado de miscigenação ao longo de séculos (PENA *et al.*, 2011). A população brasileira foi formada por processos de intensa miscigenação entre europeus, africanos e ameríndios, sendo essa mistura mais predominante no Nordeste, muito em função do

padrão de colonização que foi estabelecido desde os primórdios (RIBEIRO, 1995; MARTINS, 2007). O modelo socioeconômico da economia canavieira imposto, na época, foi determinante na composição étnica da população do NE e na sua mestiçagem (RIBEIRO, 1995).

Pernambuco é uma das nove unidades federativas que compõe a Região Nordeste (NE). Tem como limites os estados da Paraíba e Ceará (N), oceano atlântico (L), Alagoas e Bahia (S) e Piauí (O). Possui uma área total estimada em 98.067,881 km², na qual 9.616.621 de pessoas residem (IBGE, 2019). Composto por 185 municípios, o Estado está dividido em cinco mesorregiões, definidas pelo Instituto Brasileiro de Geografia e Estatística, com base em critérios socioeconômicos: Região Metropolitana do Recife, Zona da Mata, Agreste, Sertão e São Francisco (IBGE, 1990–). De acordo com a Pesquisa Nacional por Amostra de Domicílios Contínua do IBGE, baseada em autodeclaração, a população do Nordeste é composta por 60,4% de pardos, 31,2% brancos, 7,2% de pretos (IBGE, 2020).

Diferentes marcadores genéticos têm sido utilizados para conhecer a constituição genética do povo brasileiro. Já foi demonstrado que as linhagens paternas brasileiras são majoritariamente de origem europeia, enquanto as matrilineagens são distribuídas entre as origens europeia (39%), africana (28%) e ameríndia [33%] (ALVES-SILVA *et al.*, 2000; CARVALHO-SILVA *et al.*, 2001; RESQUE *et al.*, 2016). A constituição genômica da população brasileira difere entre as regiões: a hegemonia da contribuição europeia varia de cerca de 51% no Norte, a 77% no sul do país (MOURA *et al.*, 2015).

Tratando-se especificamente do Nordeste, Resque *et al.* (2016) demonstrou que as linhagens paternas dessa região são cerca de 84% de origem europeia, ~15% de origem africana e ~1% de origem ameríndia. Em contrapartida, segundo Alves-Silva *et al.* (2000), as matrilineagens encontram-se distribuídas entre as seguintes contribuições: 44% africanas, 34% europeias e 22% ameríndias. Além disso, segundo um estudo realizado por Moura *et al.*, (2015), utilizando como base publicações de pesquisas com marcadores autossômicos (período de 1980 a 2014), foi estimado que a ancestralidade genômica da população nordestina é formada pelas seguintes contribuições: ~58% europeia, ~27% africana e ~15% ameríndia (MOURA *et al.*, 2015).

Ainda assim, é possível verificar que, de um modo geral, a população brasileira guarda uma certa homogeneidade genética, com o predomínio da influência europeia em todas as regiões brasileiras (PENA *et al.*, 2011; MOURA *et al.*, 2015). Algumas peculiaridades, entretanto, são observáveis: há uma maior contribuição europeia na população da região Sul, enquanto há um forte componente de ancestralidade ameríndia, no Norte, e africana, no Nordeste (MOURA *et al.*, 2015).

Até o presente, poucos estudos com marcadores forenses foram realizados na população de Pernambuco. Dellalibera *et al.* (2004) determinaram as frequências alélicas e os parâmetros forenses de 13 STRs autossômicas do CODIS. Posteriormente, Santos *et al.* (2015) fizeram uma atualização dos parâmetros populacionais para esse mesmo conjunto de marcadores.

No que concerne aos Y-STRs, constam, no YHRD, apenas 30 “haplótipos mínimos”. Recentemente, um estudo com populações brasileiras estudou o desempenho forense de um conjunto de 27 Y-STRs, incluindo a população de Pernambuco na análise. Por meio desse trabalho, foi realizada uma deposição de 141 haplótipos no modelo “YFiler Plus” de indivíduos da cidade de Recife (JANNUZZI *et al.*, 2020). Apesar dessa importante inclusão, o estudo não relacionou indivíduos nascidos em outras mesorregiões do Estado.

Destarte, o presente trabalho teve o intuito de desenvolver bancos de dados populacionais, para STRs autossômicas e de STRs do cromossomo Y, específicos e representativos para a população do Estado de Pernambuco.

3 OBJETIVOS

3.1 GERAL

Avaliar, por meio de marcadores microssatélites autossômicos (STRs) e do cromossomo Y (Y-STRs), a diversidade genética da população do estado de Pernambuco.

3.2 ESPECÍFICOS

- Determinar as frequências alélicas dos 23 marcadores microssatélites autossômicos (STRs), sendo 20 do CODIS (CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 e D22S1045), e três auxiliares (SE33, PENTA D e PENTA E);
- Estimar as frequências alélicas e haplotípicas de 23 marcadores microssatélites cromossomo Y (Y-STRs): DYS456, DYS389I, DYS448, DYS389II, DYS19, DYS391, DYS481, DYS459, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385a/b, DYS456 e YGATAH4;
- Verificar se os marcadores STRs autossômicos estão em equilíbrio de Hardy-Weinberg (EHW);
- Analisar o desempenho forense de todas as STRs autossômicas, por meio da estimativa dos parâmetros estatísticos populacionais relacionados, tais como: heterozigosidade observada ($H_{Obs.}$), heterozigosidade esperada ($H_{Esp.}$), poder de discriminação (PD), poder de exclusão (PE), probabilidade de matching (PM), conteúdo de informação polimórfica (CIP) e índice de paternidade típico (IPTípico);
- Analisar o desempenho forense dos 23 Y-STRs, mediante parâmetros estatísticos populacionais relativos a esses marcadores: diversidade haplotípica (DH), capacidade de discriminação (CD), diversidade gênica (DG) e probabilidade de matching haplotípico (PMH);

- Avaliar, com base nos Y-STRs, o grau de diversidade genética da população estudada em relação a outras populações do Brasil e do mundo, por meio da diferenciação genética (F_{ST} e R_{ST});
- Submeter os haplótipos ao YHRD, permitindo a utilização desse banco dados para investigações forenses na população estudada.

4 ARTIGO 1 – GENETIC CHARACTERIZATION OF 23 Y CHROMOSOME SHORT TANDEM REPEATS IN THE STATE OF PERNAMBUCO, BRAZIL

Revista: *American Journal of Human Biology*

Qualis CAPES: A1

Fator de impacto: 1,937

Formato: *Original Article*

Comprovante de submissão: ANEXO B

Genetic characterization of 23 Y chromosome short tandem repeats in the State of Pernambuco, Brazil

Bruno Sampaio^{*,a}, Abigail Marcelino dos Santos Silva^a, Bárbara Natieli Silva Pereira^a, Aparecida Jayane Sampaio Miranda^a, Bruno Almeida Silva^a, Ellen Cássia de Oliveira Medeiros^a, Carlos Antônio de Souza^a, Anna Theresa de Souza Liberal^a, Sérgio de Sá Leitão Paiva Júnior^a, Marcos da Silveira Regueira Neto^a, Valdir de Queiroz Balbino^a

*Corresponding author

E-mail addresses: brunosmpaio@gmail.com (B. Sampaio), abigail.marcelino@gmail.com (A.M.S. Silva), barbara.natieli@gmail.com (B.N.S. Pereira), jayane_miranda@hotmail.com (A.J.S. Miranda), almeidasilva998@gmail.com (B.A. Silva), ellen.olivm@gmail.com (E.C.O. Medeiros), carlos.souza@sds.pe.gov.br (C.A. Souza), theresaliberal@gmail.com (A.T.S. Liberal), sslpaiva@gmail.com (S.S.L. Paiva Júnior), mareg1988@gmail.com (M.S. Regueira Neto), valdir.balbino@ufpe.br (V.Q. Balbino)

^aLaboratory of Bioinformatics and Evolutionary Biology Laboratory, Department of Genetics, Federal University of Pernambuco, Recife, Pernambuco, Brazil

Address: 1235 Professor Moraes Rego Avenue, Recife, Pernambuco (PE), 50670-901, Brazil

Abstract

Objectives: In this study, 437 healthy adult unrelated individuals, born and residents in the State of Pernambuco (Brazil) were genotyped for 23 Y-STRs (DYS576, DYS389 I, DYS448, DYS389 II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385, DYS456, YGATAH4) with the multiplex kit PowerPlex® Y23.

Methods: Intrapopulation forensic parameters were estimated for the all 23 Y-STRs. Genetic distance tests (F_{ST} and R_{ST}) were performed comparing Pernambuco and other populations. **Results:** Of the total of 437 samples, we observed 417 different haplotypes — 399 was unique, 16 shared between two individuals, and two shared between three individuals. A highly level of haplotype diversity was observed ($DH=0.9998$). Discrimination capacity and matching probability was 0.9542 and 0.00023, respectively. Genetic distance analysis (F_{ST} and R_{ST}) showed a genetic proximity of the Pernambuco and other Brazilian populations. Also revealed that the population of Pernambuco was closer to European than African and Native American populations. **Conclusions:** This panel of 23 Y-STRs showed high discriminatory power in this population. The genetic distance results were consistent with the history of Brazilian colonization. These results may be useful to forensic investigations, kinship analysis and historical investigations purposes.

Keywords: Forensic Genetics; Y chromosome short tandem repeats (Y-STRs); Pernambuco State; Brazil

1. Introduction

Brazil is a country of continental dimensions and was the cradle for the formation of one of the most heterogeneous populations in the world (Resque et al., 2016; Moura et al., 2015; Palha et al., 2012; Pena et al., 2010; Carvalho-Silva et al., 2001; Alves-Silva et al., 2000), because of five centuries of admixture between three mainly ethnic groups — Amerindians, Europeans, and sub-Saharan Africans (Pena et al., 2010; IBGE, 2007). However, this population formation process in the vast Brazilian territory varied according to distinct histories of colonization and settlement among the five geopolitical macro-regions of Brazil (North, Northeast, Central-West, Southeast and South), reflecting in the genetic diversity existing of the Brazilian population (Resque et al., 2016; IBGE, 2007).

Pernambuco is one of the nine federal units of the Northeast Region of Brazil (Supplementary Figure 1A). Its total area comprises 98,067,881 Km² and its current estimated population is about 9,674,793 inhabitants (IBGE, 2021). The state is divided into five mesoregions (Supplementary Figure 1B), defined by the Brazilian Institute of

Geography and Statistics (IBGE), based on socioeconomic and geographic criteria: Recife Metropolitan Region; Zona da Mata; Agreste; Sertao; and Sao Francisco (IBGE, 1990–). According to the last IBGE's National Census (2010), the population of Pernambuco is composed (self-declared information) of 55.3% of *Pardos* (brown people), 36.7% of Whites, 6.5% of Blacks, 0.9% of Yellows and 0.6% of Indigenous (IBGE, 2010).

Pernambuco was officially founded in 1537 (Costa, 1952) under the system of Hereditary Captaincies (IBGE, 2007). This territory became one of the two that initially were prosperous (Ribeiro, 1995), due to the sugarcane industry, highly lucrative at the time. It has made the region a target of territorial disputes for control of the manufacturing economy when the region was under Dutch regimen (1524–1530), but the number of Dutch colonizers is unknown (Carvalho-Silva et al., 2001; IBGE, 2007). The socioeconomic model based for a long time on the sugarcane industry was a key factor in the ethnic composition of this population and its miscegenation (Ribeiro, 1995).

Few forensic studies characterized specifically the population of Pernambuco regarding forensic markers. These studies comprise three investigations about autosomal short tandem repeats (STRs) (Santos et al., 2015; de Andrade et al., 2009; Dellalibera et al., 2004). Regarding the analysis Y-chromosomal short tandem repeats (Y-STRs), a total of 30 minimal haplotypes were deposited in Y-Chromosome STR Haplotype Reference Database (YHRD), under accession number YA003710. Recently, Jannuzzi et al. (2020) published a study about 12 populations of different Brazilian states, which included 141 Yfiler™ Plus haplotypes (YHRD accession number: YA004379). The authors did not investigate the populational of all state, considering that they only included individuals from Recife (capital of Pernambuco) in the study.

The analysis of Y-STRs haplotypes provides valuable information for forensic purposes, especially in sexual assault investigations, and for the kinship testing (Kayser, 2017). But these genetic markers are also useful for assessing population substructure or tracing population differentiation by testing pairwise genetic distances (Zhou et al., 2020). Therefore, to improve the understanding about the genetic diversity

of Y-STRs and enrich the database for the population of Pernambuco in the YHRD, the present study aimed to characterize, through a representative and randomized sampling, the Pernambuco population from the genetic polymorphisms of 23 Y-STRs loci perspective.

2. Material and methods

2.1 Ethical considerations and sample collection

This study was approved by the Committee for Ethics in Research on Human Beings of the Federal University of Pernambuco, through the protocol number 2.625.726 (CAAE 84801518.1.0000.5208). All samples were collected after a written informed consent obtained from all volunteers, following the ethical recommendations of this committee. Biological samples from a total of 437 healthy unrelated adult male individuals, born and residing in Pernambuco, were strategically collected in the five mesoregions: Recife Metropolitan Region (N=196); Zona da Mata (N=69); Agreste (N=84); Sertao (N=56) and Sao Francisco (N=32), corresponding to approximately ~0.005% of the population of each mesoregion.

2.2 PCR amplification

Blood samples were stored on FTA® (Whatman™ Bioscience) or NUCLEIC-CARD™ (Copan) cards. A total of 23 Y-chromosomal STRs (DYS576, DYS389 I, DYS448, DYS389 II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385, DYS456, YGATAH4) were amplified by PCR using the multiplex kit PowerPlex® Y23 System (PPY23, Promega Corporation) on a Veriti® Thermal Cycler (Thermo Fisher Scientific), following the manufacturer's recommendations. The 2800 M DNA (Promega Corporation) was used as typing control.

2.3 Genotyping

The amplified fragments were separated and detected by capillary electrophoresis on the Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific). Runs were performed in 36 cm capillaries (Applied Biosystems, Thermo Fisher Scientific) filled with POP-4™ polymer (Applied Biosystems, Thermo Fisher

Scientific). Raw data were analyzed using the GeneMapper® ID-X v.1.4 software (Thermo Fisher Scientific) to obtain the haplotypes, following the recommendations of the International Society for Forensic Genetics (ISFG) (Gusmão et al., 2006).

2.4 Statistical analysis

Intrapopulation forensic parameters were calculated for all samples and all 23 Y-STRs markers amplified by the PPY23 kit using an in-house software. Haplotype diversity (HD) was calculated as $HD = \frac{n}{n-1} (1 - \sum p_i^2)$, where n is the number of samples and p_i is the frequency of the i -th haplotype (Nei, 1987). Gene diversity (GD) was calculated for each loci using an equation like HD, from the allele frequencies of each Y-STR marker (Nei, 1987). Discrimination capacity (DC) was calculated as the ratio between the number of different haplotypes found in the sampling and the total number of samples analyzed (Purps et al., 2014). The match probability (MP) was calculated as $MP = 1 - HD$ (Peny et al., 2020).

The pairwise F_{ST} genetic distances and the corresponding non-differentiation P -values were calculated (considering 10,000 permutations) using the Arlequin v3.5.1.2 software (Excoffier and Lischer, 2010). The first pairwise F_{ST} genetic distance test was performed for the five mesoregions of Pernambuco. The second pairwise F_{ST} genetic distance test was carried out between the Pernambuco and other Brazilian and foreign population. This second analysis was illustrated on a multidimensional scale (MDS) using Vegan package v.2.5-6 (Oksanen et al., 2020). Finally, an additional analysis of genetic distance (R_{ST}) and the generation of the MDS plots was performed using the AMOVA&MDS online tools available at YHRD (<https://yhrd.org/amova>).

3 Results and discussion

3.1 Forensic parameters

The list of the PPY23 haplotypes and general information about the geographical origin of the individuals are listed in Supplementary Table 1. The forensic parameters for the set of 23 Y-STR markers obtained from a sample of 437 individuals from the Pernambuco are presented in Supplementary Table 2. Considering only the PPY23 panel, were identified 417 different haplotypes, 399 of which were unique ($n=1$),

16 haplotypes were found being shared between two individuals ($n=2$), and two haplotypes were found being shared between three individuals ($n=3$) (Supplementary Table 2). The values of HD and DC were 0.9998 and 0.9542, respectively. The probability (MP) that two randomly selected individuals from the population of Pernambuco share the same PPY23 haplotype was equal to 0.00023. Comparing these results with those obtained from Purps et al. (2014) for the Latin America population, the HD and DC values were similar, while MP value obtained in the present study was lower.

To evaluate and compare the usefulness of the PPY23 set in the studied population, the forensic parameters were also calculated for the other haplotypes sets: minimal haplotype (9 loci), PowerPlex® Y (12 loci) and Yfiler (17 loci). The PPY23 panel showed the best values for HD, DC, and MP, being therefore the most informative set. This result was expected, since the addition of the number of loci in an analysis increases the resolution capacity of a set of markers to discriminate paternal lineages (Purps et al., 2014; Hanson and Ballantyne, 2007).

The average gene was 0.6879, ranging from 0.5380 (DYS393) to 0.9237 (DYS385). Data of the allelic frequencies distribution and the gene diversities (GD) are shown in Supplementary Table 3. The GD results were similar to those obtained by Purps et al. (2014), also considering the Latin American population, in which the DYS393 marker showed the lowest GD value (0.5095), while the DYS385 multicopy locus presented, as expected, the highest GD (0.9323).

3.2 Allelic variants

Allelic variants (e.g., null allele, microvariants and duplications) were identified in some markers (Supplementary Table 4). The null allele was observed at DYS533 loci (sample HPE116), and it was confirmed by two new amplifications. A null allele in this locus have already been observed in the Brazilian population of the Sao Paulo state, detected with the Yfiler™ Plus kit (Jannuzzi et al., 2020). Duplications were found for samples HPE367 (DYS389I*28-29), HPE379 (DYS448*19-20) and HPE141 (which showed duplications in two regions — DYS389I*13-14 and DYS439*12-13). All these observed duplications are present in markers situated in azoospermia factor (AZF) segments — AZFa (DYS389I, DYS389II and DYS439) and AZFc (DYS448).

These regions are known for perform intrachromosomal rearrangements, which are important processes involved in the male fertility, and what could explain the identified duplications (Turrina et al., 2015; Bosch and Jobling, 2003). Microvariant alleles were observed only in DYS385 (11.2 and 12.2) and DYS458 (16.2, 17.2, 18.2, 19.2 and 20.2) loci. The microvariant allele 11.2 at DYS385 has not been previously observed in the most current studies involving Brazilian populations (Jannuzzi et al., 2020; Ambrósio et al., 2020; Santos et al., 2019; Purps et al., 2014; Rêgo et al., 2014; Palha et al., 2012). All duplications and microvariants were confirmed by one more amplification for each sample.

3.3 Population comparisons

The pairwise F_{ST} genetic distance analysis performed comparing the populations of the five mesoregions (Recife Metropolitan Region, Zona da Mata, Agreste, Sertao and Sao Francisco) of the Pernambuco state. The F_{ST} values were low and the non-differentiation probabilities were above the significance level (P -value ≤ 0.05), showing no statistically significant differences among these populations, and consequently revealing a genetic homogeneity of these tested populations (the results are detailed in Supplementary Table 5).

The pairwise F_{ST} genetic distance tests between Pernambuco and other Brazilian and foreign populations utilized as reference included a total of 7,097 haplotypes of PPY23 [21,29,31]. All information related to the populations used in the analysis, as well as the sample size, ancestry data and the references corresponding to each population, are described in the Supplementary Table 6. The genetic distances between all population pairs as well as the corresponding non-differentiation P -values are presented in Supplementary Table 7. The corresponding MDS graph of this analysis is shown in Supplementary Figure 2.

The results showed that the population of Pernambuco is genetically close to the Brazilian population from the Sao Paulo state ($F_{ST} = 0.0001$; P -value = 0.00306) and the population of the North of Portugal ($F_{ST} = 0.0026$; P -value = 0.0046). On the other hand, the Pernambuco population is genetically distant from the Xhosa, an African population from South Africa used as a reference for this ancestry ($F_{ST} =$

0.14987; P -value = 0.00000). The F_{ST} results were considered statistically significant at non-differentiation P -value ≤ 0.00024 , after the Bonferroni correction.

The R_{ST} genetic distances analysis was similarly performed, comparing a total of 9,109 PPY23 haplotypes (populations included in this test are shown in Supplementary Table 8). The pairwise R_{ST} genetic distances (Supplementary Table 9) was better to calculate that the proximity between Pernambuco and other Brazilian populations. In this analysis, The R_{ST} results were low and the non-differentiation probabilities were above the significance level (P -value ≤ 0.00017). It was possible to observe that this population is close to some European and South America populations analyzed.

These populations formed a cluster in the representation in MDS (Supplementary Figure 3). A MDS graph was generated exclusively to show to show the Brazilian populations (Supplementary Figure 4). On the opposite, the Pernambuco population is genetically distant (with statistically significant differences) from the Native American populations of Sao Gabriel da Cachoeira (municipality located in the state of Amazonas, Brazil) and Bolivia. The Pernambuco population was also distant from the African populations from South Africa (Xhosa) and Zimbabwe.

Our findings are in accordance with other Y-STRs previous studies, which revealed that the Brazilian populations are genetically closer to Europeans than to Africans and Native Americans (Jannuzzi et al., 2020; Ambrósio et al., 2020; Santos et al., 2019; Purps et al., 2014; Rêgo et al., 2014; Palha et al., 2012). The Brazilian paternal ancestry is majority European for all the macro-regions, already demonstrated by Y-chromosomal Single Nucleotide Polymorphisms (Y-SNPs) studies (Resque et al., 2016; Carvalho-Silva et al., 2001); on the other hand, the maternal ancestry (mtDNA) is distributed in European, African, and Native American contributions, being the African mtDNA lineages most prominent in the Northeast region (Alves-Silva et al., 2000).

These results of the present study are consistent with the colonization history of Brazil. The formation of the current Brazilian population began in 1500 with the meeting of Europeans with the Native American population composed by 2.4 million of inhabitants) at the time (Carvalho-Silva et al., 2001). The initial Portuguese

colonization of Brazil (1500–1808) involved about 500.000 Europeans, majority men (Carvalho-Silva et al., 2001). Soon after, there was an influx of about 3.6 million sub-Saharan Africans who were brought to Brazil during the slave trade period (Palha et al., 2012). During the 19th and 20th centuries, Brazil received about 6 million of immigrants from Italy, Portugal, Spain, Germany, Syria, Lebanon, and Japan, contributing genetically in the formation of the Brazilian population, despite this influx was more restricted to the Southeast and South regions of Brazil (Carvalho-Silva et al., 2001). Although the miscegenation processes occurred in Brazil, even as in some others Latin American populations (Salzano and Sans, 2014), the Brazilian patrilineages are predominantly of European origin. The male-specific portion of the Y chromosome helps to explain historical aspects in the formation of the Brazilian population, which received influences from the diverse socioeconomic models in the past, including of the sugarcane industry — established in Pernambuco during a long time (IBGE, 2007; Ribeiro, 1995).

4 Conclusion

In summary, our results showed that the populations from the five mesoregions of Pernambuco are paternally close. This enables the enrichment of the Y-STR database with 437 PPY23 haplotypes for all the Pernambuco population. Our findings corroborate with other studies with these markers realized in the Brazilian populations, reinforcing a genetic homogeneity referring to the Y chromosome ancestry. These data will contribute to forensic investigations and kinship analysis and will be useful to direct future historical and evolutionary investigations in this population.

Acknowledgments

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES, Ministry of Education, Brazil), for funding this work (Edital Pró-Forenses 25/2014). The authors also thank the volunteers who participated in this study.

References

- Alves-Silva J, da Silva Santos M, Guimarães PE, et al. The ancestry of Brazilian mtDNA lineages [published correction appears in *Am J Hum Genet* 2000 Sep;67(3):775]. *Am J Hum Genet*. 2000;67(2):444-461. doi:10.1086/303004
- Ambrosio IB, Braganholi DF, Orlando LBM, et al. Mutational data and population profiling of 23 Y-STRs in three Brazilian populations. *Forensic Sci Int Genet*. 2020;48:102348. doi:10.1016/j.fsigen.2020.102348
- Bosch E, Jobling MA. Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. *Hum Mol Genet*. 2003;12(3):341-347. doi:10.1093/hmg/ddg031
- Carvalho-Silva DR, Santos FR, Rocha J, Pena SD. The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet*. 2001;68(1):281-286. doi:10.1086/316931
- Costa FAP (1952). *Anais Pernambucanos: 1493-1590. Vol. I. Recife, Arquivo Público Estadual, 1951. 644 p.*
- de Andrade ES, Gomes AV, Raposo G, Mauricio-da-Silva L, Silva Rdos S. Mutation rates at 14 STR loci in the population from Pernambuco, Northeast Brazil. *Forensic Sci Int Genet*. 2009;3(4):e141-e143. doi:10.1016/j.fsigen.2008.11.011
- Dellalibera E, Havro ML, Souza M, Kajihara K, Mauricio-da-Silva L, dos Santos Silva R. Genetic analysis of 13 STR loci in the population from the State of Pernambuco, northeast Brazil. *Forensic Sci Int*. 2004;146(1):57-59. doi:10.1016/j.forsciint.2004.04.005
- Divisão do Brasil em mesorregiões e microrregiões geográficas. IBGE, Departamento de Geografia. 1990—.
- Excoffier, L. and H.E. L. Lischer (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*. 10: 564-567.
- Gusmão L, Butler JM, Carracedo A, et al. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs

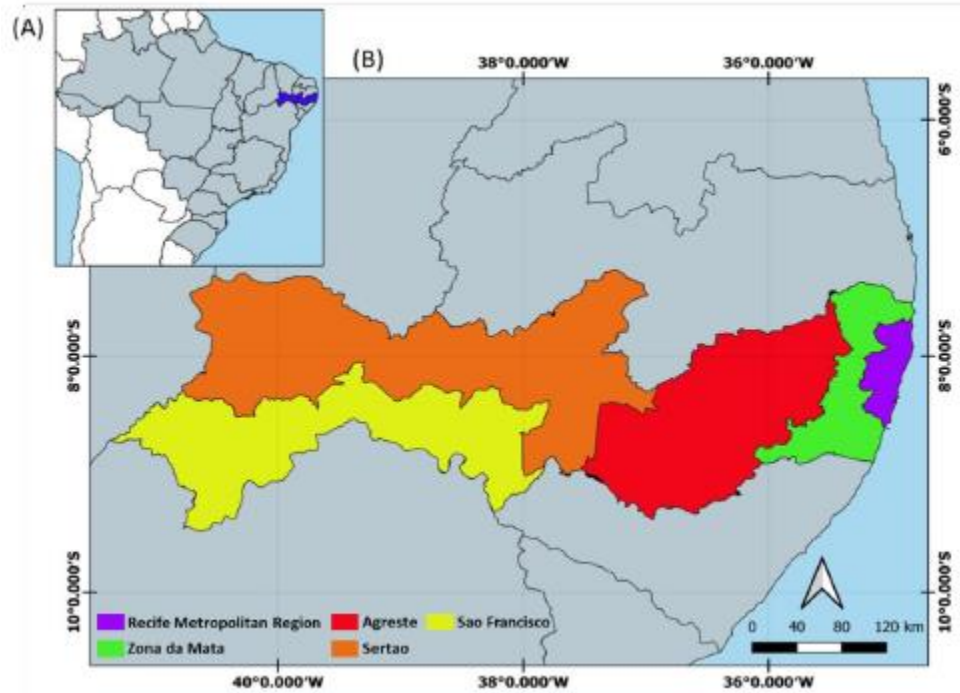
- 301 in forensic analysis. *Int J Legal Med.* 2006;120(4):191-200. doi:10.1007/s00414-005-
302 0026-1
- 303 Hanson EK, Ballantyne J. An ultra-high discrimination Y chromosome short tandem
304 repeat multiplex DNA typing system. *PLoS One.* 2007;2(8):e688. Published 2007 Aug
305 1. doi:10.1371/journal.pone.0000688
- 306 Instituto Brasileiro de Estatística (IBGE). 2007. Brasil: 500 anos de povoamento. Rio
307 de Janeiro: IBGE. p 237.
- 308 Instituto Brasileiro de Geografia e Estatística (IBGE). Censo Demográfico 2010.
- 309 Instituto Brasileiro de Geografia e Estatística (IBGE). 2021.
310 <https://www.ibge.gov.br/cidades-e-estados/pe.html>
- 311 Jannuzzi J, Ribeiro J, Alho C, et al. Male lineages in Brazilian populations and
312 performance of haplogroup prediction tools. *Forensic Sci Int Genet.* 2020;44:102163.
313 doi:10.1016/j.fsigen.2019.102163
- 314 Kayser M. Forensic use of Y-chromosome DNA: a general overview. *Hum Genet.*
315 2017;136(5):621-635. doi:10.1007/s00439-017-1776-9
- 316 Moura RR, Coelho AV, Balbino Vde Q, Crovella S, Brandão LA. Meta-analysis of
317 Brazilian genetic admixture and comparison with other Latin America countries. *Am J*
318 *Hum Biol.* 2015;27(5):674-680. doi:10.1002/ajhb.22714
- 319 Nei, M. *Molecular Evolutionary Genetics.* Columbia University Press, New York, 1987.
- 320 Oksanen, JF, Blanchet G, Friendly M et al. *vegan: Community Ecology Package, R*
321 *Package Version 2.5-7, 2020, <http://CRAN.Rproject.org/package=vegan>.*
- 322 Palha T, Gusmão L, Ribeiro-Rodrigues E, Guerreiro JF, Ribeiro-Dos-Santos A, Santos
323 S. Disclosing the genetic structure of Brazil through analysis of male lineages with
324 highly discriminating haplotypes. *PLoS One.* 2012;7(7):e40007.
325 doi:10.1371/journal.pone.0040007
- 326 Pena SD, Di Pietro G, Fuchshuber-Moraes M, et al. The genomic ancestry of
327 individuals from different geographical regions of Brazil is more uniform than expected.

- 328 *PLoS One.* 2011;6(2):e17063. Published 2011 Feb 16.
329 doi:10.1371/journal.pone.0017063
- 330 Peng D, Zhang Y, Ren H, et al. Identification of sequence polymorphisms at 58 STRs
331 and 94 iiSNPs in a Tibetan population using massively parallel sequencing. *Scientific*
332 *Reports – Nature*, v.10, p.12225, 2020. <https://doi.org/10.1038/s41598-020-69137-1>
- 333 Purps J, Siegert S, Willuweit S, et al. A global analysis of Y-chromosomal haplotype
334 diversity for 23 STR loci. *Forensic Sci Int Genet.* 2014;12(100):12-23.
335 doi:10.1016/j.fsigen.2014.04.008
- 336 Rêgo GE, Monte SM, De Moura-Neto RS, De Oliveira NF. Allele frequencies and
337 population data for 17 Y-STR loci in Paraíba population, Brazil. *Forensic Sci Int Genet.*
338 2014;13:e18-e19. doi:10.1016/j.fsigen.2014.06.002
- 339 Resque R, Gusmão L, Geppert M, et al. Male Lineages in Brazil: Intercontinental
340 Admixture and Stratification of the European Background. *PLoS One.*
341 2016;11(4):e0152573. Published 2016 Apr 5. doi:10.1371/journal.pone.0152573
- 342 Ribeiro J, Romero M, Simão F, et al. Analysis of 23 Y-STRs in a population sample
343 from eastern Paraguay. *Forensic Sci Int Genet.* 2018;37:e20-e22.
344 doi:10.1016/j.fsigen.2018.08.007
- 345 Ribeiro D (1995) *O povo brasileiro: a formação e o sentido do Brasil.* São Paulo:
346 Companhia das Letras. 476 p.
- 347 Salzano FM, Sans M. Interethnic admixture and the evolution of Latin American
348 populations. *Genet. Mol. Biol.* 2014; 37: 151-170. doi.org/10.1590/S1415-
349 47572014000200003
- 350 Santos Stange V, Silva Dos Reis R, Mariano Garcia de Souza Rodrigues F, et al.
351 Stratification among European descent and admixed Brazilian populations of Espírito
352 Santo for 27 Y-STRs. *Forensic Sci Int Genet.* 2019;41:e20-e22.
353 doi:10.1016/j.fsigen.2019.03.019
- 354 Santos SM, Souza CA, Rabelo KC, et al. Distribution of forensic marker allelic
355 frequencies in Pernambuco, Northeastern Brazil. *Genet Mol Res.* 2015;14(2):4303-
356 4310. Published 2015 Apr 30. doi:10.4238/2015.April.30.2

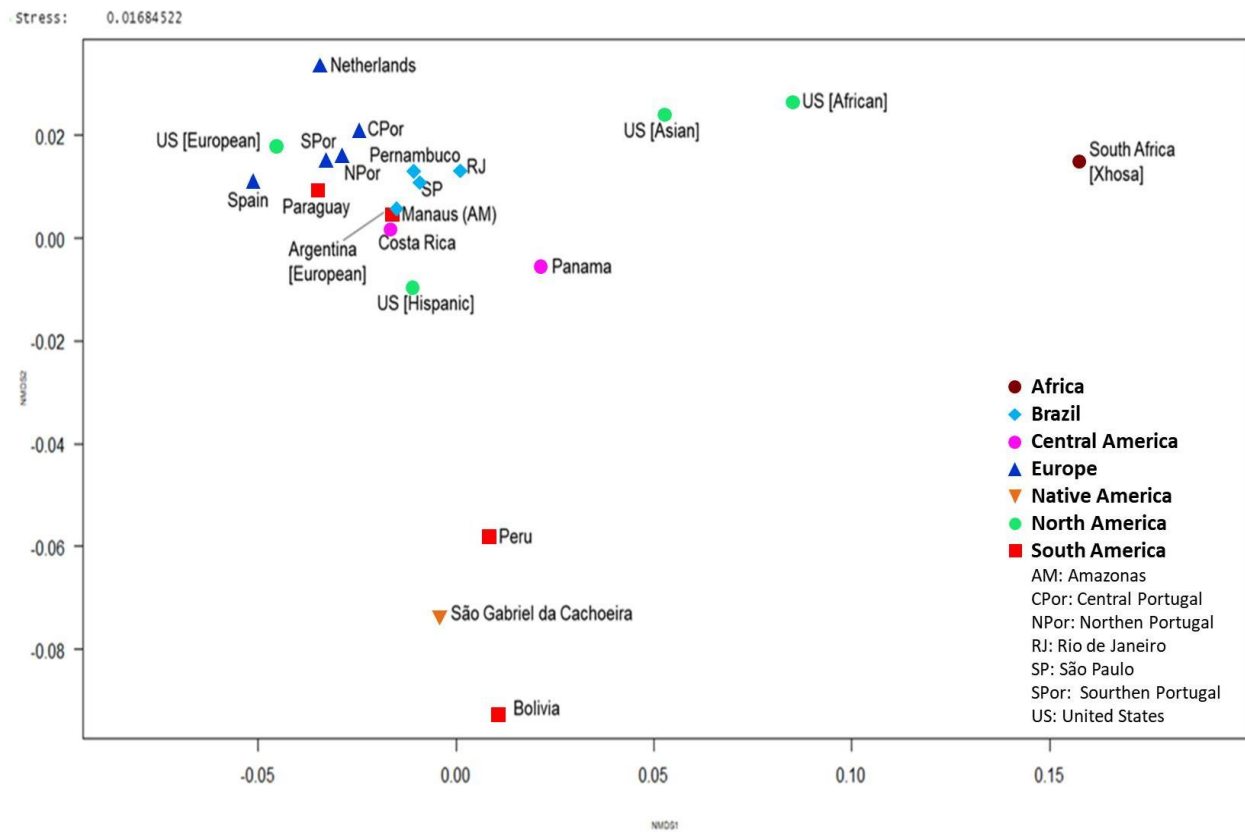
Turrina S, Caratti S, Ferrian M, De Leo D. Deletion and duplication at DYS448 and DYS626 loci: unexpected patterns within the AZFc region of the Y-chromosome. *Int J Legal Med.* 2015;129(3):449-455. doi:10.1007/s00414-015-1178-2

Zhou Y, Yao Y, Liu B, et al. Characterizing Y-STRs in the Evaluation of Population Differentiation Using the Mean of Allele Frequency Difference between Populations. *Genes (Basel).* 2020;11(5):566. Published 2020 May 19. doi:10.3390/genes11050566

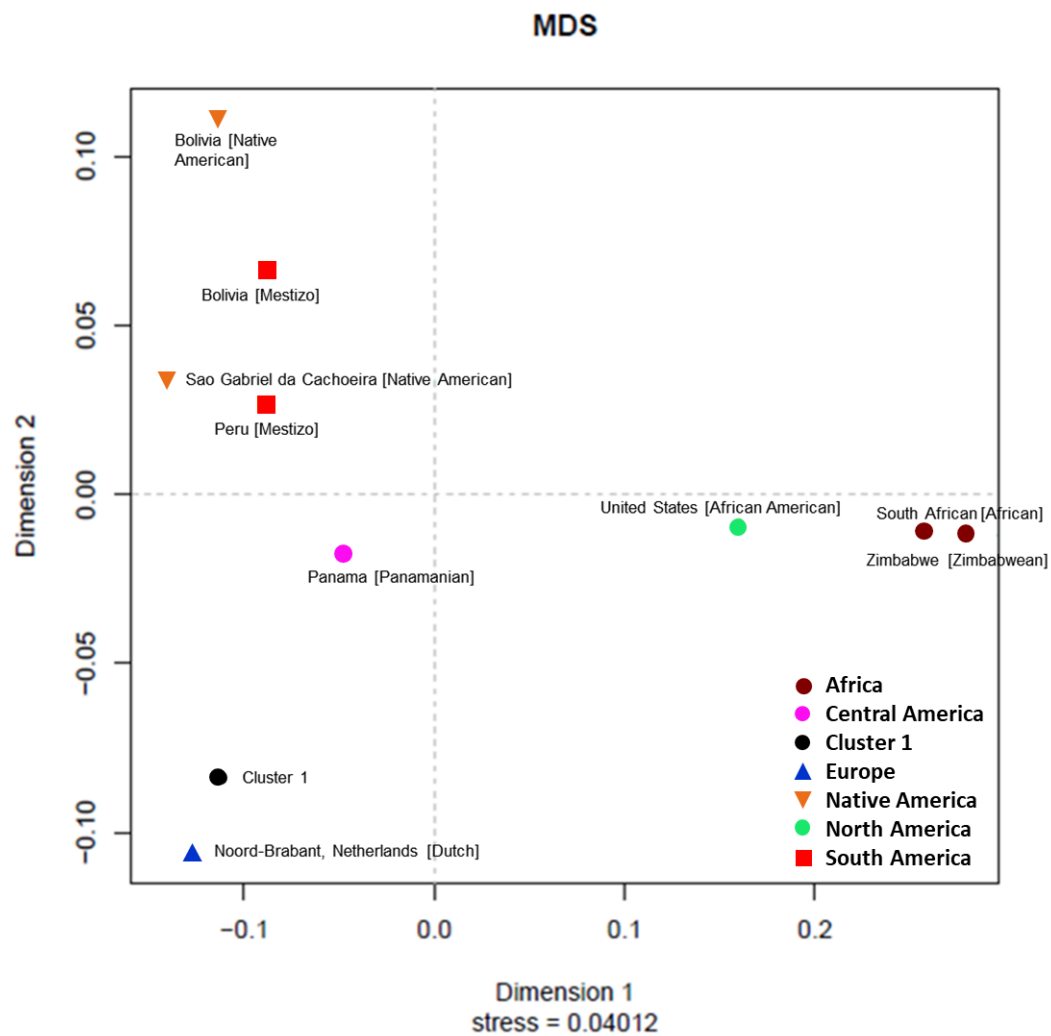
Supplementary Figure 1 – (A) Pernambuco State, located in the Brazil's Northeast Region. (B) The five mesoregions of Pernambuco: Recife Metropolitan Region, Zona da Mata, Agreste, Sertao and Sao Francisco.



Supplementary Figure 2 – Multidimensional scale (MDS) plot of the pairwise F_{ST} genetic distances, calculated between Pernambuco and other populations (Stress: 0.01685522).

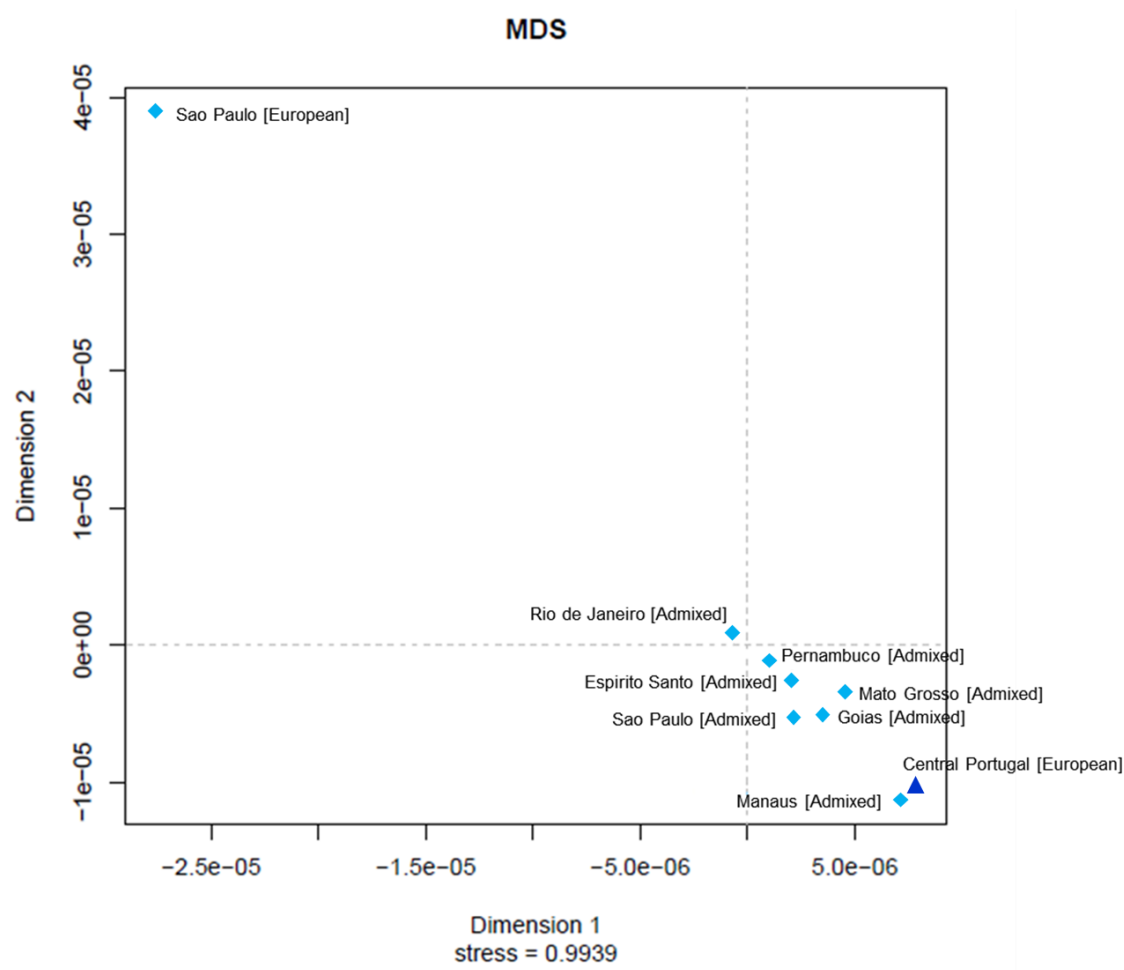


Supplementary Figure 3 – Multidimensional scale (MDS) plot of the pairwise R_{ST} genetic distances, calculated between Pernambuco and other populations.



Cluster 1: Buenos Aires, Argentina [European]; Pernambuco [Admixed]; Manaus, Brazil [Admixed Brazilian]; Goias, Brazil [Admixed]; Mato Grosso, Brazil [Admixed]; Rio de Janeiro, Brazil [Admixed]; Sao Paulo, Brazil [Admixed]; Espirito Santo, Brazil [Brazilian]; Sao Paulo, Brazil [European]; Costa Rica [Mestizo]; Friesland, Netherlands [Dutch]; Netherlands [Dutch]; Eastern Paraguay, Paraguay [Admixed]; Central Portugal, Portugal [Portuguese]; Northern Portugal, Portugal [Portuguese]; Southern Portugal, Portugal [Portuguese]

Supplementary Figure 4 – Multidimensional scale (MDS) plot of the pairwise R_{ST} genetic distances, calculated between Brazilian and Central Portugal populations.



Supplementary Table 1 – Comparison about number of haplotypes and forensic parameters considering four set of Y-STRs markers.

	Minimal	PowerPlex Y	Yfiler	PowerPlex Y23
Haplotypes (samples)	437	437	437	437
Different haplotypes	317	358	399	417
Unique (n=1)	258	311	370	399
n=2	38	33	23	16
n=3	9	10	4	2
n=4	4	1	1	-
n=5	5	-	1	-
n=6	1	1	-	-
n=7	1	1	-	-
n=13	-	1	-	-
n=22	1	-	-	-
Haplotype diversity (HD)	0.9957	0.9981	0.9995	0.9998
Discrimination capacity (DC)	0.7254	0.8192	0.9130	0.9542
Match probability (MP)	0.0043 (4.3×10^{-3})	0.0019 (1.9×10^{-3})	0.00054 (5.4×10^{-4})	0.00023 (2.3×10^{-4})
Gene diversity (GD)	0.6401	0.6608	0.6815	0.6879

Allele	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS456	YGATAH4	DYS385	
0									0.0023													9, 12	0.0023
7																	0.0023					10, 11	0.0046
8									0.0046	0.0023							0.0183					10, 13	0.0046
9						0.0664		0.0092	0.0252	0.0870							0.0984					10, 14	0.0046
10						0.4668		0.0206	0.0481	0.3181					0.0778	0.0023	0.4485				0.0297	10, 15	0.0092
11		0.0114				0.4371		0.1556	0.2975	0.1648					0.3661	0.4256	0.1167	0.0069		0.0023	0.3707	11, 11	0.0092
12		0.1968			0.0023	0.0275		0.4279	0.4874	0.4050					0.4027	0.0595	0.2151	0.1648		0.0046	0.5355	11, 12	0.0137
12, 13															0.0023							11, 13	0.0458
13	0.0069	0.6293			0.1281	0.0023		0.3227	0.1121	0.0206	0.0046				0.1350	0.4394	0.0686	0.6384	0.0023	0.0503	0.0618	11, 14	0.2380
13, 14		0.0023																				11, 15	0.0526
14	0.0137	0.1579			0.5034			0.0549	0.0229	0.0023	0.4073	0.0023			0.0137	0.0641	0.0275	0.167	0.0092	0.0938	0.0023	11, 16	0.0069
15	0.0664	0.0023			0.2471			0.0092			0.4645	0.0114			0.0023	0.0092	0.0046	0.0229	0.1487	0.4783		11, 17	0.0023
16	0.1327				0.0961						0.1167	0.1190							0.2197	0.2677		11, 18	0.0046
16.2																			0.0092			12, 12	0.0275
17	0.2494		0.0023		0.0229						0.0069	0.3547	0.0092						0.3410	0.0870		12, 13	0.0114
17.2																			0.0092			12, 14	0.0481
18	0.3204		0.0686								0.2128	0.0023							0.1808	0.0069		12, 15	0.0137
18.2																			0.0206			12, 16	0.0137
19	0.1579		0.4233				0.0046					0.1350	0.0046						0.0343	0.0069		12, 17	0.0046
19.2																			0.0092			12, 19	0.0046
19, 20			0.0023																			12, 20	0.0023
20	0.0435		0.3272				0.0114					0.0709	0.0709	0.0023					0.0137	0.0023		13, 13	0.0206
20.2																			0.0023			13, 14	0.0526
21	0.0092		0.1442				0.0801					0.0435	0.2449	0.0870								13, 15	0.0389
22			0.0206				0.3959					0.0366	0.1281	0.0938								13, 16	0.0160
23			0.0114				0.1602					0.0137	0.4050	0.2723								13, 17	0.0366
24							0.0938						0.1144	0.4371								13, 18	0.0389
25							0.0778						0.0160	0.0938								13, 19	0.0137
26				0.0092			0.0595						0.0046	0.0137								14, 14	0.0252
27				0.0229			0.0458															14, 15	0.0252
28				0.1076			0.0526															14, 16	0.0092

28, 29	0.0023																				14, 17	0.0069
29	0.4577						0.0160														14, 18	0.0023
30	0.2746						0.0023														14, 19	0.0046
31	0.1098																				14, 21	0.0023
32	0.0137																				15, 15	0.0297
33	0.0023																				15, 16	0.0206
																					15, 17	0.0114
																					15, 18	0.0092
																					15, 19	0.0046
																					15, 20	0.0046
																					15, 21	0.0023
																					16, 16	0.0229
																					16, 17	0.0183
																					16, 18	0.0229
																					17, 17	0.0092
																					17, 18	0.0160
																					17, 19	0.0023
																					17, 20	0.0023
																					18, 20	0.0023
																					19, 19	0.0023
																					19, 20	0.0023

N_Samples	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437	437
N_Different Alleles	9	6	8	9	6	5	12	7	8	7	5	10	10	7	7	6	9	5	13	10	5	52
Gene Diversity (DG)	0.7878	0.5414	0.6892	0.6923	0.6608	0.5872	0.7893	0.6865	0.6594	0.7012	0.6060	0.7897	0.7428	0.7111	0.6808	0.6195	0.7251	0.5380	0.7803	0.6821	0.5724	0.9237

Supplementary Table 3 – Allelic variants observed in the present study.

Locus	Occurrence	Allelic variants	Sample	Number of variant types per locus
DYS389I	1	13, 14 (duplication)	HPE141	1
DYS439	1	12, 13 (duplication)	HPE141	1
DYS448	1	19, 20 (duplication)	HPE367	1
DYS389II	1	28, 29 (duplication)	HPE367	1
DYS533	1	Null	HPE116	1
DYS458	20	16.2; 17.2; 18.2; 19.2; 20.2 (microvariant)	-	5
DYS385a/b	3	11.2; 12.2 (microvariant)	-	2

Supplementary Table 4 – F_{ST} genetic distances (under the diagonal) between populations of the five mesoregions of Pernambuco state and non-differentiation correspondent P -values (0.005) (above the diagonal), considering 10,000 permutations.

	Pernambuco Agreste	Pernambuco Sao Francisco	Pernambuco Zona da Mata	Pernambuco Sertao	Recife Metropolitan Region
Pernambuco Agreste	*	0.85515	0.56835	0.70965	0.66405
Pernambuco Sao Francisco	-0.00023	*	0.50375	0.52220	0.26695
Pernambuco Zona da Mata	0.00019	0.00023	*	0.73215	0.77210
Pernambuco Sertao	0.00004	0.00033	0.00002	*	0.37535
Recife Metropolitan Region	0.00001	0.00011	0.00003	0.00033	*

Supplementary Table 5 – List of included populations in the F_{ST} analysis of the present study.

Population	Ancestry	Sample size	References
Brazil			
Pernambuco	Admixed	437	Present study
Manaus, Amazonas	Admixed	204	[a]
Sao Gabriel da Cachoeira, Amazonas	Native American	61	[b]
Rio de Janeiro	Admixed	201	[a]
Sao Paulo	Admixed	200	[a]
North America			
United States	African American	509	[b]
United States	Asian American	96	[b]
United States	European American	534	[b]
United States	Hispanic American	403	[b]
South America			
Bolivia	Admixed	44	[b]
Argentina	European	264	[b]
Peru	Admixed	83	[b]
Paraguay	Admixed	537	[c]
Central America			
Panama	Admixed	100	[b]
Costa Rica	Admixed	166	[b]
Europe			
Spain	European	660	[b]
Central Portugal	European	83	[b]
Northern Portugal	European	85	[b]
Southern Portugal	European	80	[b]
Netherlands	European	2180	[b]
Africa			
South Africa (Xhosa)	African	114	[b]

References:

[a] Ambrosio IB, Braganholi DF, Orlando LBM, et al. Mutational data and population profiling of

23 Y-STRs in three Brazilian populations. *Forensic Sci Int Genet.* 2020;48:102348. doi:10.1016/j.fsigen.2020.102348

[b] Purps J, Siegert S, Willuweit S, et al. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Sci Int Genet.* 2014;12(100):12-23. doi:10.1016/j.fsigen.2014.04.008

[c] Ribeiro J, Romero M, Simão F, et al. Analysis of 23 Y-STRs in a population sample from eastern Paraguay. *Forensic Sci Int Genet.* 2018;37:e20-e22. doi:10.1016/j.fsigen.2018.08.007

Supplementary Table 6 – F_{ST} genetic distances (under the diagonal) between Pernambuco population and other reported populations and non-differentiation correspondent P -values (0.00024, after Bonferroni's correction) (above the diagonal), considering 10,000 permutations.

	Pernambuco	Manaus, Amazonas	Sao Gabriel de Cachoeira, Brazil [Native American]	Rio de Janeiro	Sao Paulo	United States [African American]	United States [Asian American]	United States [European American]	United States [Hispanic American]	Bolivia	Argentina [European]	Peru	Paraguay	Panama	Costa Rica	Spain	Central Portugal	Northern Portugal	Southern Portugal	Netherlands	South Africa [Xhosa]
Pernambuco	*	0.00000	0.01345	0.00000	0.00306	0.00000	0.00206	0.00000	0.00000	0.02751	0.00000	0.00053	0.00000	0.00000	0.00000	0.00000	0.00023	0.00460	0.00000	0.00000	0.00000
Manaus, Amazonas	0.00216	*	0.20092	0.20850	0.47282	0.15409	0.52167	0.14620	0.06079	0.50068	0.00000	0.00905	0.00000	0.54419	0.00000	0.00177	0.02858	0.55109	0.01483	0.00000	0.00000
Sao Gabriel de Cachoeira, Brazil [Native American]	0.06977	0.05598	*	0.15406	0.21097	0.32170	0.13674	0.36758	0.17933	0.06162	0.00036	0.00183	0.00000	0.14602	0.00000	0.03545	0.00106	0.14621	0.00008	0.00000	0.00000
Rio de Janeiro	0.00289	0.00561	0.06864	*	0.53678	0.13197	0.59383	0.17555	0.06447	0.51986	0.00000	0.00790	0.00000	0.42141	0.00000	0.00000	0.23750	0.57426	0.00905	0.00000	0.00000
Sao Paulo	0.0001	0.00366	0.06757	0.00169	*	0.46016	1.00000	0.43427	0.04724	0.63058	0.00000	0.01555	0.00000	1.00000	0.00055	0.01257	0.02607	1.00000	0.01704	0.00000	0.00000
United States [African American]	0.06188	0.07074	0.11686	0.04982	0.05476	*	0.31871	0.12202	0.00000	0.58383	0.00000	0.05715	0.00000	0.34353	0.00000	0.00000	0.12377	0.43188	0.04717	0.00000	0.00000
United States [Asian American]	0.06858	0.07757	0.12268	0.06254	0.07228	0.07469	*	0.57755	0.26881	0.56204	0.00000	0.00405	0.00000	1.00000	0.00064	0.03826	0.01662	1.00000	0.01535	0.00000	0.00000
United States [European American]	0.01309	0.01406	0.09188	0.01974	0.01584	0.10967	0.1149	*	0.01413	0.58711	0.00000	0.10966	0.00000	0.66402	0.00000	0.00000	0.07057	0.44350	0.08299	0.00000	0.00000
United States [Hispanic American]	0.00531	0.00300	0.04713	0.00693	0.00406	0.06461	0.06636	0.01914	*	0.39614	0.00000	0.03592	0.00000	0.26523	0.00000	0.00000	0.04930	0.25563	0.04078	0.00000	0.00619
Bolivia	0.07623	0.07023	0.06905	0.07864	0.07745	0.11562	0.12113	0.10632	0.05291	*	0.00484	0.00266	0.00000	0.58434	0.00624	0.42447	0.00924	0.53309	0.00602	0.01424	0.00064
Argentina [European]	0.00168	0.0028	0.05969	0.00375	0.00198	0.06857	0.0658	0.01273	0.00309	0.07247	*	0.00000	0.00000	0.00000	0.00000	0.00000	0.00023	0.00009	0.00000	0.00000	0.00000
Peru	0.03825	0.03361	0.05196	0.04195	0.03741	0.08608	0.09144	0.05878	0.02253	0.01904	0.03134	*	0.00000	0.00133	0.00000	0.00104	0.00000	0.00055	0.00000	0.00000	0.00000
Paraguay	0.00626	0.00747	0.07544	0.01156	0.00687	0.09344	0.09368	0.0049	0.00899	0.08651	0.00509	0.04468	*	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
Panama	0.01109	0.0122	0.05484	0.00878	0.00857	0.03261	0.05331	0.0423	0.00902	0.05231	0.01131	0.02772	0.02719	*	0.00298	0.01466	0.01147	1.00000	0.00636	0.00000	0.00000
Costa Rica	0.00301	0.00313	0.06127	0.00458	0.00195	0.06947	0.07461	0.00987	0.00184	0.0688	0.00162	0.03054	0.00448	0.01249	*	0.00000	0.00000	0.00011	0.00000	0.00000	0.00000
Spain	0.01388	0.01459	0.08923	0.02196	0.01438	0.11285	0.11357	0.00516	0.0158	0.10421	0.01193	0.05742	0.00246	0.04131	0.01119	*	0.01585	0.01786	0.00799	0.00000	0.00000
Central Portugal	0.00358	0.00575	0.08508	0.00968	0.00417	0.08056	0.07796	0.01372	0.0094	0.09638	0.0054	0.04743	0.00467	0.02148	0.00717	0.01087	*	0.02228	0.00010	0.00000	0.00000
Northern Portugal	0.0026	0.00294	0.08308	0.00696	0.00301	0.08948	0.0924	0.00377	0.00939	0.09256	0.00369	0.04967	0.00135	0.02477	0.00273	0.00433	0.0027	*	0.02779	0.00000	0.00000
Southern Portugal	0.00492	0.00545	0.08433	0.0083	0.00437	0.089	0.09838	0.00385	0.00961	0.0966	0.00344	0.04943	-0.00065	0.02523	0.00378	0.00038	0.00248	-0.00222	*	0.00000	0.00000
Netherlands	0.0152	0.01644	0.09532	0.01992	0.01931	0.10466	0.10797	0.00465	0.02264	0.11307	0.01426	0.06399	0.01269	0.04291	0.01398	0.01365	0.01726	0.00738	0.00981	*	0.00000
South Africa [Xhosa]	0.14987	0.16168	0.20402	0.13418	0.14777	0.0507	0.12644	0.20987	0.14593	0.19844	0.15725	0.17505	0.1893	0.11269	0.16205	0.21207	0.1832	0.20166	0.19721	0.19647	*

Supplementary Table 7 – List of included populations in the R_{ST} analysis of the present study.

Population	Ancestry*	Sample size *	YHRD accession number/References
Brazil			
Pernambuco [Admixed]	Admixed	408	Present study
Manaus [Admixed Brazilian]	Admixed Brazilian	201	YA004319 [a]
Goiás [Admixed]	Admixed	67	YA004577
Mato Grosso [Admixed]	Admixed	190	YA004248
Rio de Janeiro [Admixed]	Admixed	385	YA003720, YA004282, YA004619 [a]
Sao Paulo [Admixed]	Admixed	481	YA004247, YA004620 [a], YA004685
Sao Paulo [European]	European American	119	YA003142 [b]
Espirito Santo [Brazilian]	Admixed	32	YA003849
Sao Gabriel da Cachoeira [Native American]	Native American	60	YA003788
North America			
United States [African American]	African American	2449	YA004417, YA004422, YA004455
South America			
Bolivia [Native American]	Native American	55	YA003827
Bolivia [Mestizo]	Admixed	44	YA003826
Buenos Aires, Argentina	European American	359	YA002987 [c]
Peru [Mestizo]	Admixed	80	YA003399 [d]
Paraguay	Admixed	511	YA004326 [e]
Central America			
Panama [Panamanian]	Admixed	92	YA003888
Costa Rica [Mestizo]	Admixed	160	YA003353
Europe			
Central Portugal [Portuguese]	European	80	YA003016 [f]
Northern Portugal [Portuguese]	European	84	YA003015 [g]
Southern Portugal [Portuguese]	European	78	YA003017 [g]
Netherlands [Dutch]	European	2061	YA002897 [h], YA004301 [i]
Noord-Brabant, Netherlands [Dutch]	European	14	YA003653 [i]
Friesland, Netherlands [Dutch]	European	93	YA003881
Africa			
Zimbabwe [Zimbabwean]	African	525	YA003884, YA004732
South Africa [African]	African	182	YA004588 [j]

*Sample size after exclusions of variants by YHRD

References:

- [a] Ambrosio IB, Braganholi DF, Orlando LBM, et al. Mutational data and population profiling of 23 Y-STRs in three Brazilian populations. *Forensic Sci Int Genet.* 2020;48:102348. doi:10.1016/j.fsigen.2020.102348
- [b] Gusmão L, Sánchez-Diz P, Calafell F, et al. Mutation rates at Y chromosome specific microsatellites. *Hum Mutat.* 2005;26(6):520-528. doi:10.1002/humu.20254
- [c] Corach D, Filgueira Risso L, Marino M, Penacino G, Sala A. Routine Y-STR typing in forensic casework. *Forensic Sci Int.* 2001;118(2-3):131-135. doi:10.1016/s0379-0738(00)00483-7
- [d] Iannaccone GC, Tito RY, Lopez PW, Medina ME, Lizarraga B. Y-chromosomal haplotypes for the PowerPlex Y for twelve STRs in a Peruvian population sample. *J Forensic Sci.* 2005;50(1):239-242.
- [e] Ribeiro J, Romero M, Simão F, et al. Analysis of 23 Y-STRs in a population sample from eastern Paraguay. *Forensic Sci Int Genet.* 2018;37:e20-e22. doi:10.1016/j.fsigen.2018.08.007
- [f] Bento AM, Carvalho M, Lopes V, et al. Distribution of Y-chromosomal haplotypes in the Central Portuguese population using 17-STRs. *Forensic Sci Int Genet.* 2009;4(1):e35-e36. doi:10.1016/j.fsigen.2009.02.006
- [g] Alves C, Gomes V, Prata MJ, Amorim A, Gusmão L. Population data for Y-chromosome haplotypes defined by 17 STRs (AmpFISTR YFiler) in Portugal. *Forensic Sci Int.* 2007;171(2-3):250-255. doi:10.1016/j.forsciint.2006.10.022
- [h] Westen AA, Kraaijenbrink T, Clarisse L, et al. Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males. *Forensic Sci Int Genet.* 2015;14:174-181. doi:10.1016/j.fsigen.2014.10.012
- [i] Claerhout S, Roelens J, Van der Haegen M, Verstraete P, Larmuseau MHD, Decorte R. Ysurnames? The patrilineal Y-chromosome and surname correlation for DNA kinship research. *Forensic Sci Int Genet.* 2020;44:102204. doi:10.1016/j.fsigen.2019.102204
- [j] Reid KM, Heathfield LJ. Allele frequency data for 23 Y-chromosome short tandem repeats (STRs) for the South African population. *Forensic Sci Int Genet.* 2020;46:102270. doi:10.1016/j.fsigen.2020.102270

Supplementary Table 8 – R_{ST} genetic distances between (under the diagonal) between Pernambuco population and other reported populations and non-differentiation correspondent P -values (0.00017, after Bonferroni's correction) (above the diagonal), considering 10,000 permutations.

	Pernambuco [Admixed]	South Africa [African]	Buenos Aires, Argentina [European]	Bolivia [Mestizo]	Bolivia [Native American]	Manaus, Amazonas [Admixed Brazilian]	Goiás, Brazil [Admixed]	Mato Grosso, Brazil [Admixed]	Rio de Janeiro, Brazil [Admixed]	Sao Paulo, Brazil [Admixed]	Espirito Santo, Brazil [Brazilian]	Sao Paulo, Brazil [European]	Sao Gabriel da Cachoeira, Brazil [Native]	Costa Rica [Mestizo]	Friesland, Netherlands [Dutch]	Netherlands [Dutch]	Noord-Brabant, Netherlands [Dutch]	Panama [Panamanian]	Eastern Paraguay, Paraguay [Admixed]	Peru [Mestizo]	Central Portugal, Portugal [Portuguese]	Northern Portugal, Portugal [Portuguese]	Southern Portugal, Portugal [Portuguese]	United States [African American]	Zimbabwe [Zimbabwean]
Pernambuco [Admixed]	*	0.0000	0.0023	0.0000	0.0000	0.0305	0.8333	0.1764	0.6743	0.1180	0.7681	0.2501	0.0000	0.1061	0.0082	0.0000	0.0567	0.0001	0.0000	0.0000	0.2038	0.0053	0.0218	0.0000	0.0000
South Africa [African]	0.3545	*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003
Buenos Aires, Argentina [European]	0.0072	0.3774	*	0.0000	0.0000	0.1763	0.2777	0.0334	0.0002	0.0105	0.3145	0.0106	0.0000	0.0562	0.0001	0.0000	0.0628	0.0000	0.0003	0.0000	0.0604	0.0004	0.0097	0.0000	0.0000
Bolivia [Mestizo]	0.1378	0.3903	0.1025	*	0.0132	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0243	0.0000	0.0000	0.0000	0.0000	0.0000
Bolivia [Native American]	0.2088	0.3959	0.1706	0.0279	*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Manaus, Brazil [Admixed Brazilian]	0.0049	0.3997	0.0015	0.1244	0.2053	*	0.6298	0.2081	0.0037	0.0185	0.2640	0.0179	0.0000	0.3225	0.0016	0.0000	0.1363	0.0000	0.0662	0.0000	0.5777	0.0237	0.1263	0.0000	0.0000
Goiás, Brazil [Admixed]	-0.0041	0.3697	0.0014	0.1302	0.2251	-0.0026	*	0.7736	0.5641	0.4825	0.5639	0.2977	0.0000	0.6409	0.0841	0.0209	0.1465	0.0078	0.1064	0.0001	0.3965	0.0427	0.0957	0.0000	0.0000
Mato Grosso, Brazil [Admixed]	0.0015	0.3759	0.0052	0.1315	0.2125	0.0015	-0.0043	*	0.0651	0.2785	0.4614	0.0387	0.0000	0.3313	0.0090	0.0000	0.1213	0.0002	0.0177	0.0000	0.2844	0.0385	0.1847	0.0000	0.0000
Rio de Janeiro, Brazil [Admixed]	-0.0008	0.3317	0.0098	0.1379	0.2041	0.0097	-0.0021	0.0040	*	0.1287	0.8291	0.3732	0.0000	0.0459	0.0060	0.0000	0.0449	0.0005	0.0000	0.0000	0.0505	0.0016	0.0078	0.0000	0.0000
Sao Paulo, Brazil [Admixed]	0.0013	0.3446	0.0046	0.1255	0.1916	0.0060	-0.0012	0.0006	0.0014	*	0.8238	0.1122	0.0000	0.0839	0.0003	0.0000	0.0430	0.0004	0.0000	0.0000	0.0566	0.0012	0.0130	0.0000	0.0000
Espirito Santo, Brazil [Brazilian]	-0.0066	0.3313	0.0016	0.1464	0.2281	0.0034	-0.0057	-0.0024	-0.0079	-0.0077	*	0.3555	0.0002	0.2100	0.0445	0.0147	0.0699	0.1058	0.0390	0.0013	0.1694	0.0212	0.1466	0.0000	0.0000
Sao Paulo, Brazil [European]	0.0011	0.3280	0.0103	0.1287	0.2059	0.0112	0.0015	0.0089	0.0000	0.0035	0.0005	*	0.0000	0.0665	0.0032	0.0000	0.0166	0.0091	0.0001	0.0001	0.0953	0.0005	0.0036	0.0000	0.0000
Sao Gabriel da Cachoeira, Brazil [Native American]	0.1125	0.4172	0.0791	0.0754	0.1105	0.0918	0.1222	0.1132	0.1128	0.0995	0.1199	0.1047	*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
Costa Rica [Mestizo]	0.0031	0.3844	0.0042	0.1277	0.2086	0.0005	-0.0031	0.0003	0.0055	0.0034	0.0058	0.0072	0.1042	*	0.0087	0.0000	0.0986	0.0003	0.0071	0.0000	0.1930	0.0098	0.0324	0.0000	0.0000
Friesland, Netherlands [Dutch]	0.0141	0.4259	0.0289	0.2164	0.3089	0.0218	0.0109	0.0182	0.0171	0.0240	0.0268	0.0283	0.1855	0.0195	*	0.6195	0.1444	0.0000	0.0015	0.0000	0.0187	0.0395	0.0204	0.0000	0.0000
Netherlands [Dutch]	0.0204	0.4855	0.0358	0.2335	0.3141	0.0227	0.0148	0.0227	0.0261	0.0328	0.0329	0.0411	0.1881	0.0229	-0.0017	*	0.1610	0.0000	0.0000	0.0000	0.0025	0.0184	0.0098	0.0000	0.0000
Noord-Brabant, Netherlands [Dutch]	0.0361	0.4844	0.0310	0.1916	0.3216	0.0179	0.0233	0.0223	0.0436	0.0429	0.0472	0.0662	0.1881	0.0267	0.0205	0.0168	*	0.0025	0.2199	0.0007	0.1051	0.2885	0.2999	0.0000	0.0000
Panama [Panamanian]	0.0352	0.2565	0.0344	0.0884	0.1428	0.0470	0.0304	0.0396	0.0280	0.0270	0.0137	0.0209	0.0975	0.0415	0.0806	0.1101	0.1027	*	0.0000	0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
Eastern Paraguay, Paraguay [Admixed]	0.0125	0.4473	0.0095	0.1552	0.2431	0.0031	0.0057	0.0059	0.0205	0.0161	0.0195	0.0276	0.1266	0.0093	0.0198	0.0194	0.0097	0.0804	*	0.0000	0.3956	0.0730	0.1722	0.0000	0.0000
Peru [Mestizo]	0.0618	0.3654	0.0352	0.0223	0.0770	0.0514	0.0550	0.0586	0.0630	0.0505	0.0590	0.0559	0.0471	0.0491	0.1252	0.1358	0.1154	0.0412	0.0768	*	0.0000	0.0000	0.0000	0.0000	0.0000
Central Portugal, Portugal [Portuguese]	0.0024	0.4110	0.0071	0.1570	0.2589	-0.0017	-0.0004	0.0012	0.0088	0.0078	0.0097	0.0076	0.1324	0.0035	0.0190	0.0205	0.0278	0.0565	-0.0001	0.0757	*	0.3510	0.2222	0.0000	0.0000
Northern Portugal, Portugal [Portuguese]	0.0177	0.4597	0.0254	0.2107	0.3184	0.0113	0.0163	0.0115	0.0250	0.0251	0.0325	0.0357	0.1816	0.0183	0.0132	0.0119	0.0058	0.0939	0.0056	0.1206	0.0007	*	0.4006	0.0000	0.0000
Southern Portugal, Portugal [Portuguese]	0.0120	0.4370	0.0149	0.1851	0.2829	0.0047	0.0110	0.0032	0.0181	0.0147	0.0106	0.0293	0.1463	0.0134	0.0189	0.0163	0.0053	0.0737	0.0030	0.0979	0.0036	-0.0002	*	0.0000	0.0000
United States [African American]	0.1931	0.0549	0.2156	0.2509	0.2612	0.2312	0.1974	0.2064	0.1760	0.1828	0.1622	0.1698	0.2596	0.2105	0.2430	0.2830	0.3131	0.1209	0.2622	0.2059	0.2388	0.2764	0.2573	*	0.0000
Zimbabwe [Zimbabwean]	0.4112	0.0135	0.4383	0.4709	0.4702	0.4628	0.4402	0.4401	0.3903	0.4019	0.4094	0.3970	0.4872	0.4466	0.4825	0.5128	0.5549	0.3379	0.4930	0.4399	0.4773	0.5163	0.5008	0.0804	*

**5 ARTIGO 2 – ALLELIC FREQUENCIES DISTRIBUTION AND FORENSIC
PARAMETERS OF 23 AUTOSOMAL SHORT TANDEM REPEATS IN THE
POPULATION OF THE STATE OF PERNAMBUCO, BRAZIL**

Revista: *Legal Medicine*

Qualis CAPES: A1

Fator de impacto: 1,376

Formato: *Short Communication*

Comprovante de submissão: ANEXO C

Allelic frequencies distribution and forensic parameters of 23 autosomal short tandem repeats in the population of the State of Pernambuco, Brazil

Bruno Sampaio^a, Abigail Marcelino dos Santos Silva^a, Sérgio de Sá Paiva Leitão Júnior^a, Anna Theresa de Souza Liberal^a, Heidi Lacerda Alves da Cruz^a, Valdir de Queiroz Balbino^a

^aLaboratory of Bioinformatics and Evolutionary Biology, Department of Genetics, Federal University of Pernambuco, Recife, Pernambuco (PE), Brazil

Address: 1235 Professor Moraes Rego Avenue, Recife, Pernambuco (PE), 50670-901, Brazil Phone: +55 81 2126-8512

Highlights

- Allele frequencies and forensic parameters for 23 STRs were determined
- The sample involved 767 individuals of the State of Pernambuco, Brazil
- The most and less informative markers were SE33 and TPOX, respectively
- This panel was very informative for forensic purposes in this population

Abstract

Located in the Northeast Region, the Pernambuco State is one of the 27 federative units of Brazil. Here, we determined the allele frequencies and the populational parameters of forensic relevance of 23 short tandem repeat markers (STRs) — CSF1PO, FGA, TH01, TPOX, vWA, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, PENTA D, PENTA E and SE33. The sample consisted of 767 healthy, adult, unrelated individuals (437 males, 330 females) born and resident in the State of Pernambuco. STRs were amplified using three multiplex kits, according to the availability: PowerPlex® Fusion 6C System (Promega

Captaincy — the first Portuguese administrative system of territorial division of Brazil [4]. The Captaincy of Pernambuco was the one of the two that initially were successful, due to the sugarcane industry, highly profitable at the time, receiving a large part of the Africans to work in the sugarcane farms of this region [5]. The territory was under Dutch regimen for 30 years (1524–1530), but the number of Dutch colonizers is unknown [2,5].

When the genetic inheritance of the Brazil is accessed by autosomal and Y chromosome polymorphic markers, is revealed a certain genetic homogeneity between the Brazilian macro-regions, being the European ancestry the more pronounced in the Brazilians [2,6–7]. However, after the European genetic contribution, the African genomic influence in the Northeast Region is greater in comparison to the other regions [6].

Short tandem repeats (STRs) — also denominated as microsatellites — are stretches of DNA in which the repeat unit is composed of 1 to 7 nucleotides in length that are consecutively repeated several times [8], covering approximately 3% of the human genome [9]. These markers are widely applied in forensic routines, also being is extremely important in paternity and kinship tests [10–11]. The technical way for typing these markers based on multiplex polymerase chain reaction (PCR) and capillary electrophoresis is still forensically considered the gold standard methodology [11].

Because of the importance of the generation of more representative populational databases of these genetic markers for forensic purposes, the aim of the of this research was establish a database for 23 autosomal STRs including, being 20 autosomal of the expanded CODIS set (CSF1PO, FGA, TH01, TPOX, vWA, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, and D22S1045) and three additional markers (PENTA E, PENTA D and SE33).

2. Material and methods

2.1 Ethical considerations, participants, and sample collection

This study was approved by the Committee for Ethics in Research on Human Beings of the Federal University of Pernambuco, through the protocol number 2.625.726 (CAAE 84801518.1.0000.5208). Biological samples (blood) were obtained from a total of 767 unrelated healthy adult individuals (437 males, 330 females) born and residents in the State of Pernambuco were stored on FTA[®] (Whatman[™] Bioscience) or NUCLEIC-CARD[™] (Copan) cards. The volunteers participated of the routine caseworks performed at the Laboratory of Bioinformatics and Evolutionary Biology, located in the Federal University of Pernambuco (Recife, Brazil), responsible for carrying out paternity and kinship tests. A written informed consent was obtained from each participant in this study.

2.2 DNA amplification

For the STRs amplification, were used three multiplex kits, according to the availability: PowerPlex[®] Fusion 6C System (Promega Corporation), PowerPlex[®] Fusion System (Promega Corporation) and GlobalFiler[™] Express (Thermo Fisher Scientific). The amplification was performed directly from 1.2 mm punches following the recommendations contained in the manufacturer's manuals, using a Veriti[®] Thermal Cycler (Thermo Fisher Scientific).

2.3 Genotyping

The amplified fragments were separated and detected by capillary electrophoresis on an Applied Biosystems 3500 Genetic Analyzer (Thermo Fisher Scientific). Runs were carried out in 36 cm capillaries (Applied Biosystems, Thermo Fisher Scientific) filled with POP-4[™] polymer (Applied Biosystems, Thermo Fisher Scientific). Raw data were analyzed by the GeneMapper[®] ID-X v.1.4 software (Thermo Fisher Scientific). Only one profile of each subject was included after deletion of replicates.

2.4 Statistical analysis

Allele frequencies, expected heterozygosity (H_{exp}), observed heterozygosity (H_{obs}), matching probability (MP), power of discrimination (PD), power of exclusion (PE), typical paternity index (TPI) and polymorphism information content (PIC) were calculated using an in-house software, developed with algorithms designed from formulas contained in the following literature [12–19]. Tests to estimate deviations from

Table 1 – Statistical data about populational parameters of forensic relevance of 23 autosomal STRs for the Pernambuco population.

Locus	N	Hobs	Hexp	MP	PD	PE	TPI	PIC	PHW
CSF1PO	765	0.7281	0.7406	0.1137	0.8863	0.4731	1.8389	0.7242	0.0074
THO1	764	0.8010	0.8018	0.0691	0.9309	0.6010	2.5132	0.7931	0.0887
TPOX	760	0.7158	0.7041	0.1304	0.8696	0.4531	1.7593	0.6974	0.0050
VWA	755	0.7974	0.8085	0.0624	0.9376	0.5941	2.4673	0.7965	0.1089
D16S539	766	0.7781	0.7844	0.0795	0.9205	0.5590	2.2529	0.7673	0.6435
D7S820	763	0.7785	0.8003	0.0662	0.9338	0.5598	2.2574	0.7857	0.0273
D13S317	766	0.7742	0.7892	0.0729	0.9271	0.5520	2.2139	0.7707	0.2813
D18S51	766	0.8786	0.8848	0.0241	0.9759	0.7519	4.1183	0.8826	0.0576
D21S11	764	0.8416	0.8512	0.0384	0.9616	0.6784	3.1570	0.8510	0.0037
D8S1179	764	0.8024	0.8130	0.0589	0.9411	0.6034	2.5298	0.7984	0.2671
D3S1358	764	0.7801	0.7703	0.0881	0.9119	0.5627	2.2738	0.7476	0.8665
PENTAE	285	0.9018	0.9083	0.0142	0.9858	0.7990	5.0893	0.9063	0.8872
D12S391	732	0.8866	0.8768	0.0274	0.9726	0.7682	4.4096	0.8766	0.3167
FGA	762	0.8871	0.8768	0.0279	0.9721	0.7692	4.4302	0.8762	0.5251
D5S818	764	0.7487	0.7458	0.1066	0.8934	0.5075	1.9896	0.7144	0.7351
PENTAD	283	0.8657	0.8716	0.0290	0.9710	0.7261	3.7237	0.8681	0.4629
D1S1656	766	0.9008	0.8931	0.0206	0.9794	0.7970	5.0395	0.8925	0.5216
D2S441	720	0.7444	0.7561	0.0991	0.9009	0.5003	1.9565	0.7553	0.0083
D2S1338	766	0.8747	0.8870	0.0225	0.9775	0.7440	3.9896	0.8852	0.1959
D10S1248	721	0.7781	0.7794	0.0825	0.9175	0.5590	2.2531	0.7574	0.2434
SE33	569	0.9192	0.9423	0.0063	0.9937	0.8347	6.1848	0.9421	0.0180
D19S433	761	0.8344	0.8283	0.0514	0.9486	0.6644	3.0198	0.8273	0.0597
D22S1045	718	0.7284	0.7477	0.0978	0.9022	0.4736	1.8410	0.7199	0.5228

- a) N: number of individuals; Hobs: observed heterozygosity; Hexp: expected heterozygosity; MP: matching probability; PD: power of discrimination; PE: power of exclusion; TPI: typical paternity index; PIC: polymorphic information content.
- b) P -value = 0.0022 (after Bonferroni correction).
- c) Highlighted in red: the most (SE33) and less (TPOX) informative markers, considering the best and worst parameters values, respectively.

4. Discussion

In the present study, we obtained the allele frequencies distribution and the populational parameters of forensic relevance of 23 short tandem repeat markers (CSF1PO, FGA, TH01, TPOX, vWA, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, PENTA D, PENTA E and SE33).

According to results of the forensic parameters, the SE33 marker showed the best values for Hobs, Hexp, PD, PE, TPI, PIC and MP. These results were similar who those found by Moysés et al. [21], considering only the Brazilian Northeast population. This study was carried out with the GlobalFiler™ Express panel and analyzed the five macro-regions Brazilian populations. Our results also indicates that the SE33 locus is, in fact, the more informative for this population. This locus is of great interest to the forensic community because of its highly polymorphic and heterozygotic nature, providing a more discriminating genotype, being especially important for DNA mixture analysis [22].

Other markers that showed best forensic performance was PENTA E and D1S1656. The first is a pentanucleotide high polymorphic STR genetic marker, which in our study was amplified with the kits PowerPlex® Fusion and PowerPlex® Fusion 6C. The second was the most informative marker of Brazil's population in the study of Aguiar et al. [23], which worked with the expanded CODIS core loci.

On the other hand, the TPOX marker showed the lowest Hobs, Hexp, PD, PE, TPI and PIC values, and the highest MP value. This finding is according to those obtained by Moysés et al. [21], considering the population of the Northeast, and by Aguiar et al. [23], considering the Brazilian population in general, reveling that it is the least informative marker for forensic purposes.

5. Conclusion

The present study evaluated the forensic utility of 23 autosomal STRs loci for the population of Pernambuco, a Northeast Brazilian State. This is the first effort to consider the data for all these markers simultaneously in a Brazilian population. The obtained results make possible to create a more current and specific database for the Pernambuco State and show that these 23 autosomal STRs loci are very informative being of high value in terms of forensic and kinship analysis for this population.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES, Ministry of Education, Brazil), for funding this work (Edital Pró-Forenses 25/2014). The authors also thank the volunteers who participated in this study.

Contributions

Bruno Sampaio: Conceptualization; acquisition, analysis, and interpretation of data; and writing original draft. Abigail Marcelino: Conceptualization; analysis, and interpretation of data; and statistical analysis. Sérgio Leitão Júnior: Conceptualization; interpretation of data; and statistical analysis. Theresa Liberal: Acquisition, analysis, and interpretation of data; Heidi Cruz: Acquisition and analysis of data. Valdir Balbino: Conceptualization; analysis, and interpretation of data; and writing original draft.

References

- [1] Pena SD, Di Pietro G, Fuchshuber-Moraes M, et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected.

- PLoS One*. 2011;6(2):e17063. Published 2011 Feb 16. doi:10.1371/journal.pone.0017063
- [2] Carvalho-Silva DR, Santos FR, Rocha J, Pena SD. The phylogeography of Brazilian Y-chromosome lineages. *Am J Hum Genet*. 2001;68(1):281-286. doi:10.1086/316931
- [3] Palha T, Gusmão L, Ribeiro-Rodrigues E, Guerreiro JF, Ribeiro-Dos-Santos A, Santos S. Disclosing the genetic structure of Brazil through analysis of male lineages with highly discriminating haplotypes. *PLoS One*. 2012;7(7):e40007. doi:10.1371/journal.pone.0040007
- [4] Instituto Brasileiro de Estatística (IBGE). 2007. Brasil: 500 anos de povoamento. Rio de Janeiro: IBGE. p 237.
- [5] Ribeiro D (1995) O povo brasileiro: a formação e o sentido do Brasil. São Paulo: Companhia das Letras. 476 p.
- [6] Moura RR, Coelho AV, Balbino Vde Q, Crovella S, Brandão LA. Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *Am J Hum Biol*. 2015;27(5):674-680. doi:10.1002/ajhb.22714
- [7] Resque R, Gusmão L, Geppert M, et al. Male Lineages in Brazil: Intercontinental Admixture and Stratification of the European Background. *PLoS One*. 2016;11(4):e0152573. Published 2016 Apr 5. doi:10.1371/journal.pone.0152573
- [8] Butler JM. Short tandem repeat typing technologies used in human identity testing. *Biotechniques*. 2007;43(4):ii-v. doi:10.2144/000112582
- [9] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome [published correction appears in *Nature* 2001 Aug 2;412(6846):565] [published correction appears in *Nature* 2001 Jun 7;411(6838):720. Szustakowski, J [corrected to Szustakowski, J]]. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062
- [10] Mo SK, Liu YC, Wang SQ, et al. Exploring the efficacy of paternity and kinship testing based on single nucleotide polymorphisms. *Forensic Sci Int Genet*. 2016;22:161-168. doi:10.1016/j.fsigen.2016.02.012

- [11] Tan Y, Bai P, Wang L, et al. Two-person DNA mixture interpretation based on a novel set of SNP-STR markers. *Forensic Sci Int Genet.* 2018;37:37-45. doi:10.1016/j.fsigen.2018.07.021
- [12] Nei, M. *Molecular Evolutionary Genetics*. Columbia University Press, New York, 1987.
- [13] Huston, KA. *Statistical Analysis of STR Data. Technical Tips, Profiles in DNA, GenePrint™*, p. 14-15, 1998.
- [14] Fisher, RA. Standard calculations for evaluating a blood-group system. *Heredity (Edinb)*, v.5, n.1, p. 95-102, 1951.
- [15] Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet.* 1980;32(3):314-331.
- [16] Guo X, Elston RC. Linkage information content of polymorphic genetic markers. *Hum Hered.* 1999;49(2):112-118. doi:10.1159/000022855
- [17] Brenner, CH & Morris, JW. Paternity Index Calculations in Single Locus Hypervariable DNA Probes: Validation and Other Studies. *The International Symposium on Human Identification*, Promega Corporation, Madison, WI, USA, p. 21-53, 1989.
- [18] Butler, J. M. *Advanced topics in Forensic DNA Typing – interpretation*. Ed.1, Academic Press, 2015.
- [19] Imad HH, Aamera JO, Abeer FM, Ghaidaa JM. Allele frequency data of 21 autosomal short tandem repeat loci in Mesan and Basra provinces in South Iraq, *Egyptian Journal of Forensic Sciences*, Volume 5, Issue 4, 2015, Pages 150-156.
- [20] Excoffier, L. and H.E. L. Lischer (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources.* 10: 564-567.
- [21] Moysés CB, Tsutsumida WM, Raimann PE, 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.* 2017;26:e28-e30. doi:10.1016/j.fsigen.2016.10.017

[22] Borsuk LA, Gettings KB, Steffen CR, Kiesler KM, Vallone PM. Sequence-based US population data for the SE33 locus. *Electrophoresis*. 2018;39(21):2694-2701. doi:10.1002/elps.201800091

[23] Aguiar VR, de Castro AM, Almeida VC, Malta FS, Ferreira AC, Louro ID. New CODIS core loci allele frequencies for 96,400 Brazilian individuals. *Forensic Sci Int Genet*. 2014;13:e6-e12. doi:10.1016/j.fsigen.2014.04.015

6 CONCLUSÕES

- O conjunto de 23 STRs autossômicos mostrou-se bastante eficiente para a população estudada: todos os *loci* analisados não apresentaram desvios em relação ao EHW e foram bastante informativos, apresentando um elevado valor para aplicações forenses e testes de parentesco nessa população.
- O conjunto de 23 Y-STRs também apresentou um elevado poder discriminatório de linhagens paternas nessa população. Além disso, foram identificadas diversas variações alélicas, como microvariantes, alelos nulos e duplicações. Todos esses achados serão úteis para os testes forenses e de parentesco relacionados a essa população.
- A análise de distância genética revelou que não há diferenças genéticas paternas entre as populações das cinco mesorregiões do de Pernambuco.
- Essa análise mostrou também que a população de Pernambuco está geneticamente próxima das populações de outros estados brasileiros (condizentes com outros estudos que revelam que há uma homogeneidade genética), estando também mais próxima das populações europeias do que de populações africanas e nativo americanas, corroborando com a história dos processos da colonização e povoamento do Brasil.
- O presente estudo vai possibilitar enriquecimento do banco de dados de Y-STRs (YHRD) para a população de Pernambuco, por meio do compartilhamento de 437 perfis haplotípicos gerados no modelo PowerPlex® Y23.
- O presente estudo irá contribuir para futuros estudos (forenses e evolutivos) a serem realizados, voltados para o Estado de Pernambuco.

REFERÊNCIAS

- ALVES-SILVA, J. *et al.* The ancestry of Brazilian mtDNA lineages. **American Journal of Human Genetics**, v.67, p. 444–461, 2000.
- ALEXANDER, R. P. *et al.* Annotating non-coding regions of the genome. **Nature Reviews Genetics**, v. 11, p. 559–571, 2010.
- ALVAREZ-CUBERO *et al.* Methodology for Y Chromosome Capture: A complete genome sequence of Y chromosome using flow cytometry, laser microdissection and magnetic streptavidin-beads. **Scientific Reports**, v. 8, n. 9436, p. 1-13, 2018.
- BALANOVSKY, O. Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. **Human Genetics**, v. 6, n. 5, p. 575-590, 2017.
- BACHTROG, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. **Nature Reviews Genetics**, v. 14, n. 2, p. 113-24, 2013.
- BALDING, D. J.; NICHOLS, R. A. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. **Forensic Science International**, v. 64, p. 125-140, 1994.
- BARBARO, Anna. Forensic Genetics. *In*: BARBARO, A. **Manual of Forensic Science: an international survey**. 1ª ed. Flórida: CRC Press, 2018. p. 89-110.
- BEIGUELMAN, Bernardo. A Lei de Hardy e Weinberg. *In*: BEIGUELMAN, B. **Genética de Populações Humanas**. 1 ed. São Paulo: Sociedade Brasileira de Genética, 2008. p. 9-28.
- BIEBER, F. R. *et al.* Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. **BMC Genetics**, v. 17, n. 125, p. 1-15, 2016.
- BUCKLETON, J. S. *et al.* Validating Databases. *In*: BUCKLETON, J. S.; BRIGHT, J.; TAYLOR, D. **Forensic DNA evidence interpretation**. 2 ed. Florida: CRC Press, 2016. p. 133-180.
- BUDOWLE, B., *et al.* CODIS and PCR-Based Short Tandem Repeat Loci: Law Enforcement Tools. *In*: BUDOWLE, B., *et al.* **Proceedings of the second European symposium on human identification**, Madison: Promega, 1998.
- BUTLER, John Marshall. **Forensic DNA typing: biology, technology and genetics of STR markers**. 2ª ed. Burlington: Elsevier Academic Press, 2005.
- BUTLER, John Marshall. **Fundamental of forensic DNA typing**. 1ª ed. Burlington: Elsevier Academic Press, 2010.
- BUTLER, J. M. Short tandem repeat typing technologies used in human identity testing. **Biotechniques**, v. 43, n. 4, p. ii-v. 2007.

BUTLER, J. M.; HILL, C. R. Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. **Forensic Science Review**, v. 24, p. 15-26, jan, 2012.

CALACAL, G. *et al.* Identification of exhumed remains of fire tragedy victims using conventional methods and autosomal/Y-chromosomal short tandem repeat DNA profiling. **American Journal of Forensic Medicine and Pathology**, v. 26, n. 3, p. 285-91, 2005.

CALLEGARI-JACQUES *et al.* Historical Genetics: Spatiotemporal Analysis of the Formation of the Brazilian Population. **American Journal of Human Biology**, v. 15, p.824–834, 2003.

CARRACEDO, Angel. 2013. Forensic Genetics: History. *In*: HOUCK, M. M. **Forensic Biology – Advanced Forensic Science Series**. 1ª ed. Londres: Elsevier Academic Press. 2015. p. 19-26.

CARRACEDO, A.; SÁNCHEZ-DIZ, P. Forensic DNA-Typing Technologies: A Review. *In*: CARRACEDO, A. **Forensic DNA Typing Protocols**. 1ª ed. Nova Jérsei: Humana Press Inc. 2005. p. 1-11.

CHARLESWORTH, B. The organization and evolution of the human Y chromosome. **Genome Biology**, v. 4, n. 9, p. 226, 2003.

CHAKRABORTY, R.; STIVERS, D. N.; SU, B.; ZHONG, Y.; BUDOWLE, B. The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. **Electrophoresis**, v. 20, n. 8, p. 1682- 96, 1999.

CARVALHO-SILVA, D. R. *et al.* The phylogeography of Brazilian Y-chromosome lineages. **American Journal of Human Genetics**, v. 68, p. 281–286, 2001.

DARCY, Ribeiro. **O povo brasileiro: a formação e o sentido do Brasil**. 2ª ed. São Paulo: Companhia das Letras, 1995.

DELLALIBERA, E. *et al.* Genetic analysis of 13 STR loci in the population from the State of Pernambuco, northeast Brazil. **Forensic Science International**, v. 146, p. 57–59, 2004.

de KNIJFF, P. *et al.* Chromosome Y microsatellites: population genetic and evolutionary aspects. **International Journal of Legal Medicine**, v. 110, p. 134- 140, 1997.

EXCOFFIER, L.; LISCHER, H. E. L. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. **Molecular Ecology Resources**, v. 10, p. 564-567, 2010.

FAN, H.; CHU, J. Y. A brief review of short tandem repeat mutation. **Genomics Proteomics Bioinformatics**, v. 5, p. 7-14, 2007.

FBI – FEDERAL BUREAU OF INVESTIGATION. Disponível em: <https://www.fbi.gov/services/laboratory/biometric-analysis/codis>. Acesso em: 2021.

FRANÇA, Genival Veloso de. **Medicina legal**. 9ª ed., Rio de Janeiro: Guanabara-Kogan, 2011.

FRANCEZ P. A. C. *et al.* Haplotype diversity of 17 Y-STR loci in an admixed population from the Brazilian Amazon. **Genetics and Molecular Biology**, v. 35, p. 45-52, 2012.

GETTINGS, K. B. *et al.* Sequence-based U.S. population data for 27 autosomal STR loci. **Forensic Science International: Genetics**, v. 37, p. 106-115, 2018.

GERESICK, H.; WIRTH, I. Genetic Kinship Investigation from Blood Groups to DNA Markers. **Transfusion Medicine and Hemotherapy**, v. 39, n. 3, p.163-175, 2012.

GODINHO, N. M. O. Banco de dados de DNA: uma ferramenta a serviço da justiça. **Revista Brasileira de Estudos de Segurança Pública**, v. 7, n. 2, p. 20- 30, 2014.

GOODWIN, W.; LINACRE, A; HADI, S. **An introduction to forensic genetics**. 2ª ed. Chichester: John Wiley & Sons Ltd., 2011.

GRIFFITHS, A J. F. *et al.* **Introdução à genética**. 11ª ed. Rio de Janeiro: Guanabara-Logan, 2016.

HANSON, E. K.; BALLANTYNE, J. An ultra-high discrimination y chromosome short tandem repeat multiplex dna typing system. **PLoS One**, v. 2, n. 8, p. 1-14, 2007.

HARES, D. R. Expanding the CODIS core loci in the United States. **Forensic Science International: Genetics**, v. 6(1), e52–54, 2012.

HARTL, D. L.; CLARK, A. G. **Princípios de Genética de Populações**. 4ª ed. Porto Alegre: Artmed Editora S.A., 2010.

HUSTON, K. A. **Statistical Analysis of STR Data. Profiles in DNA**. Madison: Promega, 1998.

JANNUZZI *et al.* Male lineages in Brazilian populations and performance of haplogroup prediction tools. **Forensic Science International: Genetics**, v. 44, p. 1–7, 2020.

JEFFREYS, A. J.; WILSON, V.; THEIN, S. L. Hypervariable ‘minisatellite’ regions in human DNA. **Nature**, v. 314, p. 67-73, 1985a.

JEFFREYS, A. J.; WILSON, V.; THEIN, S. L. Individual specific “fingerprints” of human DNA. **Nature**, v. 316, p. 76-79, 1985b.

JOBLING, M. A.; PETER, G. Encoded evidence: DNA in forensic analysis. **Nature Reviews Genetics**, v. 5, p. 739–751, 2004.

JIN, L.; ZHONG, Y.; CHAKRABORTY, R. The exact numbers of possible microsatellite motifs. **American Journal of Human Genetics.**, v. 55, n. 3, p. 582-583, 1994.

JOBLING, M.A. Curiosity in the genes: the DNA fingerprinting story. **Investigative Genetics**, v. 4, p. 20, 2013.

KAYSER, M *et al.* Evaluation of Y-chromosomal STRs: a multicenter study. **International Journal of Legal Medicine**, v. 110, p. 125-133, 1997.

KAYSER, M. Forensic use of Y-chromosome DNA: a general overview. **Human Genetics**, v. 136, p. 621-635, 2017.

KAYSER, M.; BALLANTYNE, K. N. Y Chromosome in Forensic Science. *In*: **Forensic DNA applications: an interdisciplinary perspective**. Nova York: CRC Press, 2014. p. 105-134.

KAUPPI, L. *et al.* Distinct properties of the XY pseudoautosomal region crucial for male meiosis. **Science**, v. 331, n. 6019, p. 916-920, 2011.

KOPSIDA, E. *et al.* The Role of the Y Chromosome in Brain Function. **Open Neuroendocrinology Journal**, v. 2, p. 20-30, 2009.

LARNEY, C.; BAILEY, T. L.; KOOPMAN, P. Switching on sex: transcriptional regulation of the testis-determining gene Sry. **Development**, v. 141, n. 11, p. 2195–2205, 2014.

LIEHR, T. Repetitive Elements in Humans. **International Journal of Molecular Sciences**, v. 22 (4), p. 1-10, 2021.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA ESTATÍSTICA. **Brasil: 500 Anos de Povoamento**. Rio de Janeiro: IBGE. 2007.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA ESTATÍSTICA. **Divisão do Brasil em mesorregiões e microrregiões geográficas**. Rio de Janeiro: IBGE. 1990-.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **PNADC 2020**. Disponível em: <https://sidra.ibge.gov.br/>. Acesso em: 15 ago. 2021.

MARTINS, Alexandre Magalhães. **Análise genético-histórica de haplótipos e haplogrupos do cromossomo Y humano no nordeste brasileiro**. 2007. Dissertação (Mestrado em Genética) – Universidade Federal de Pernambuco, Recife, 2007.

McMANUS, C.; PAIVA, S.; CORRÊA, P. S.; SEIXAS, L.; MELO, C.B. **Estatísticas para descrever genética de populações**. INCT: Informação genético-sanitária da pecuária brasileira, p. 1-50, 2011.

MOHAMMED, Ahmed Abdulla Ahmed. **Analysis of populations within the UAE using tandem repeat DNA markers**. 2001. Tese (Doutorado em Ciências e Medicina Forenses) – Universidade de Glasgow, Glasgow, 2001.

MOURA *et al.*, Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. **American Journal of Human Biology**, v. 27(5), 674–680, 2015.

MULERO, J. J. *et al.* Development and validation of the AmpFISTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. **Forensic Science International: Genetics**, v. 51, p. 64-75, 2006.

NAKAHORI, Y.; TAKENAKA, O.; NAKAGOME, Y. A human X-Y homologous region encodes “amelogenin”. **Genomics**, v. 9, n. 2, p. 264-269, 1991.

NAVARRO-COSTA, P. Sex, rebellion and decadence: the scandalous evolutionary history of the human Y chromosome. **Biochimica et Biophysica Acta**, v. 1822, n. 12, p. 1851-63, 2012.

NEI, M.; TAJIMA, F. DNA polymorphism detectable by restriction endonucleases. **Genetics**, v. 97, n. 1, p. 145-163, 1981.

NEI, M. **Molecular Evolutionary Genetics**. New York: Columbia University Press, 1987.

OLDT, R. F.; KANTHASWAMY, S. Expanded CODIS STR allele frequencies – Evidence for the irrelevance of race-based DNA databases. **Legal Medicine**, v. 42, p. 1-28, 2000.

OKSANEN, J.F. *et al.* vegan: Community Ecology Package, R Package Version 2.5-7, 2020. Disponível em: <http://CRAN.Rproject.org/package=vegan>. Acesso: 08 dez. 2021.

PAIVA JÚNIOR, Sérgio de Sá Leitão. **DNAstr.com: Ferramenta estatística aplicada à genética forense**. 2016. Tese (Doutorado em Genética) – Universidade Federal de Pernambuco, Recife, 2016.

PARSON, W. Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review. **Gerontology**, p. 1-7, 2018.

PENA, S. D. J. *et al.* The genomic ancestry of individuals from diferente geographical regions of Brazil is more uniform than expected. **PloS One**, v.6, p. 1-9, 2011.

PRIMORAC, D.; SCHANFIELD, M.; MARJANOVIĆ, D. Basic genetics and human genetic variation. *In*: **Forensic DNA applications: an interdisciplinary perspective**. Nova York: CRC Press, 2014. p. 3-54.

QUINTANA-MURCI, L.; FELLOUS, M. The Human Y Chromosome: The Biological Role of a "Functional Wasteland". **Journal of Biomedicine and Biotechnology**, v. 1, n. 1, p. 18-24, 2001.

RESENDE, Ana Filipa Tavares. **Estudo dos marcadores genéticos dos sistemas CODIS e ESS na população imigrante de Cabo Verde residente em Lisboa.** 2016. Dissertação (Mestrado em Medicina Legal e Ciências Forenses) – Universidade de Lisboa, Lisboa, 2016.

RIBEIRO, D. **O povo brasileiro: a formação e o sentido do Brasil.** São Paulo: Companhia das Letras. 476 p., 1995.

RISCH, N. J.; DEVLIN, B. On the probability of matching DNA fingerprints. **Science**, v. 255, p. 717-720, 1992.

ROEWER, L.; EPPLEN, J. Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work. **Forensic Science International**, v. 53, p. 163- 171, 1992.

ROEWER, L; WILLUWEIT, S. Y-chromosome short tandem repeats in forensics—Sexing, profiling, and matching male DNA. **WIREs Forensic Science**, p. 1-13, 2019.

ROEWER, L. *et al.* DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis. **Forensic Science International: Genetics**, v.48, p. 1-7, 2020.

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, p. 5463-5467, 1977.

SANTOS, S. M. *et al.* Distribution of forensic marker allelic frequencies in Pernambuco, Northeastern Brazil. **Genetics and Molecular Research**, v. 2, p. 4303–4310, 2010.

SANTOS, Sandra Maria dos. **Identificação Humana como Ferramenta de Investigações Criminais: Estudo de Frequências Alélicas de Marcadores de Interesse Forense no Estado de Pernambuco.** 2014. Tese (Doutorado em Biologia Aplicada à Saúde) – Universidade Federal de Pernambuco, Recife, 2014.

SKALETISKY, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. **Nature**, v. 423, p. 825-837, 2003.

SHEWALE, J. G. *et al.* Development and Validation of a Y-Chromosome STR Genotyping System, Y-PLEX™ 12, for Forensic Casework. **Journal of Forensic Sciences**, v. 49, n. 6, p. 1278-90, 2004.

SILVA, A. L.; DORNELAS, H. G.; CALIGIORNE, S. M.; MARINHO, P. A. Bancos de Perfis Genéticos Criminais no Brasil: Histórico e Evolução. **Brazilian Journal of Forensic Sciences, Medical Law and Bioethics**, v. 9(4), p. 499-520, 2020.

SOUTHERN, E. M. Detection of specific sequences among DNA fragments separated by gel electrophoresis. **Journal of Molecular Biology**, v. 98, p. 503-517, 1975.

SUBRAMANIAN, S.; MISHRA, R. K.; SINGH, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. **Genome Biology**, v. 4(2), R13, 2003.

TILLMAR, Andreas. **Populations and Statistics in Forensic Genetics**. 2010. Tese (Doutorado em Medicina Experimental e Clínica) – Universidade de Linköping, Linköping, 2010.

THOMAS, N. S.; HASSOLD, T. J. Aberrant recombination and the origin of Klinefelter syndrome. **Human Reproduction Update**, v. 9, n.4, p. 309-317, 2003.

THOMPSON, J. M. *et al.* Developmental validation of the PowerPlex® Y23 System: A single multiplex Y-STR analysis system for casework and database samples. **Forensic Science International: Genetics**, v. 7, n. 2, p. 240-50, 2013.

TROMBETTA, B.; CRUCIANI, F. Y chromosome palindromes and gene conversion. **Human Genetics**, v. 136, p. 605-619, 2017.

VAN OORSCHOT, R. A. H.; BALLANTYNE, K. N.; MITCHELL, R. J. Forensic trace DNA: a review. **Investigative Genetics**, v. 1, n. 14, p. 1-17, 2010.

WATERS, P. D.; WALLIS, M. C.; GRAVES, J. A. M. Mammalian sex--Origin and evolution of the Y chromosome and SRY. **Seminars in Cell & Developmental Biology**, v. 18, n. 3, p. 389-400, 2007.

WYMAN, A. R.; WHITE, R. A highly polymorphic locus in human DNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 77, n. 11, p. 6754-6758, 1980.

YHRD – Y-CHROMOSOME STR HAPLOTYPE REFERENCE DATABASE.
Disponível em: <https://yhrd.org/>. Acesso em: 10 fev. 2022.

ZERJAL, T. *et al.* The Genetic Legacy of the Mongols. **American Journal of Human Genetics**, v. 72, n. 3, p. 717-721, 2003.

**ANEXO A – COMPROVANTE DE SUBMISSÃO DO ARTIGO À REVISTA
*AMERICAN JOURNAL OF HUMAN BIOLOGY.***

American Journal of Human Biology
Original Article

Genetic characterization of 23 Y chromosome short tandem repeats in the State of Pernambuco, Brazil

Submission Status	Submitted	This submission has been sent to the editorial office and cannot be edited. Further instructions will be emailed to you from Manuscript Central. View Submission Overview
Submitted On	14 February 2022 by Bruno Sampaio	
Submission Started	14 February 2022 by Bruno Sampaio	

ANEXO B – COMPROVANTE DE SUBMISSÃO DO ARTIGO À REVISTA *LEGAL MEDICINE*.

Legal Medicine

Allelic frequencies distribution and forensic parameters of 23 autosomal short tandem repeats in the population of the State of Pernambuco, Brazil
--Manuscript Draft--

Manuscript Number:	
Article Type:	Short Communication
Keywords:	Forensic Genetics; autosomal short tandem repeats; Pernambuco population; Brazil
Corresponding Author:	Bruno Sampaio, M.Sc Federal University of Pernambuco: Universidade Federal de Pernambuco BRAZIL
First Author:	Bruno Sampaio, M.Sc
Order of Authors:	Bruno Sampaio, M.Sc
	Abigail Marcelino dos Santos Silva
	Sérgio de Sá Paiva Leitão Júnior
	Theresa de Souza Liberal
	Heidi Lacerda Alves da Cruz
	Valdir de Queiroz Balbino
Abstract:	Highlights • Allele frequencies and forensic parameters for 23 STRs were determined • The sample involved 767 individuals of the State of Pernambuco, Brazil • The most and less informative markers were SE33 and TPOX, respectively • This panel was very informative for forensic purposes in this population
Suggested Reviewers:	