UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

DEPARTAMENTO DE BIOLOGIA GERAL

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

TESE DE DOUTORADO

PERFIL GENÉTICO DA POPULAÇÃO BRASILEIRA DETERMINADO A PARTIR DE STRs (*SHORT TANDEM REPEATS*) UTILIZADOS EM APLICAÇÕES FORENSES

ORIENTADO: LAÉLIA MARIA PINTO

ORIENTADOR: EDUARDO MARTIN TARAZONA SANTOS

BELO HORIZONTE

2014

LAÉLIA MARIA PINTO

PERFIL GENÉTICO DA POPULAÇÃO BRASILEIRA DETERMINADO A PARTIR DE STRs (*SHORT TANDEM REPEATS*) UTILIZADOS EM APLICAÇÕES FORENSES

Tese de doutorado apresentada ao Programa de Pós-Graduação em Genética do Departamento de Biologia Geral do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Genética.

ORIENTADO: LAÉLIA MARIA PINTO

ORIENTADOR: EDUARDO MARTIN TARAZONA SANTOS

BELO HORIZONTE

2014

AGRADECIMENTOS

Agradeço ao meu orientador Eduardo Tarazona Santos pela oportunidade de desenvolver este trabalho, pelos ensinamentos durante todo esse tempo e por entender as minhas limitações de tempo. Agradeço, pois estes 8 anos de aprendizados foram cruciais para a minha formação profissional.

À Professora Luciana dos Santos por todo apoio nos momentos mais difíceis dessa caminhada, se não fosse ela talvez eu não chegasse até aqui. Obrigada pelo apoio, pelo carinho e pela atenção.

Ao Programa de Pós-Graduação em Genética pela oportunidade que tanto contribuiu para minha formação.

Às meninas da Biocod, Cristiane Lommez de Oliveira, Valéria Matarelli, Kátia Torres e Márcia Brant, por acreditar em mim, por me dar a possibilidade de realização deste trabalho, pelos ensinamentos, pelo carinho, pela atenção, pela paciência, enfim por tudo que vivemos durante os 10 anos de trabalho na Biocod. A vocês devo grande parte da profissional que sou hoje.

Ao Instituto Hermes Pardini por apoiar este trabalho, pelas oportunidades, pela atenção e pela disponibilidade. Em especial, agradeço a Vanessa Oliveira e Amanda Castro pelo apoio e ensinamentos.

À Carla por estar sempre ao meu lado, pelos ensinamentos que foram à base de tudo, pela amizade e carinho.

À família Biocod pela ajuda e convivência.

Aos amigos do Instituto Hermes Pardini pela colaboração. Em especial à Camila Bernardes pela ajuda no final da tese.

Aos colegas do LDGH pelo companheirismo, pela disponibilidade e por compartilhar os ensinamentos. Admiro o trabalho de vocês.

Aos meus pais, Laelio e Ana, pelo amor e apoio incondicional. Ao meu irmão, Leandro, por estar sempre ao meu e pelos conselhos sempre pertinentes. Sem o apoio de vocês nada disso seria possível.

Ao Luiz, pelo amor, paciência, companheirismo, por sempre tornar o caminho mais fácil e minha vida mais feliz.

Às minhas amigas, Elisângela, Kenia, Simone, Cristiane Aparecida, Gisele e Christiane Goveia, e as amigas da graduação, Cris, Clara, Fê e Mari, pelo carinho, por estarem sempre ao meu lado e por nossos momentos sempre tão divertidos.

À minha família pelo apoio, pelo carinho e pelos momentos felizes.

À família Souza, Lu, José Otávio e Josi, pelo apoio, carinho e acolhimento.

Aos amigos de Sabará pelo companheirismo.

A todos que me ajudaram nessa caminhada... MUITO OBRIGADA!

“Quando a gente acha que tem todas as respostas,

vem a vida e muda todas as perguntas”

Luis Fernando Veríssimo

ÍNDICE

LISTA DE FIGURAS

Capítulo I

Capítulo II

LISTA DE TABELAS

Capítulo I

Capítulo II

LISTA DE ABREVIATURAS

STR	*Short Tandem Repeats* - repetições curtas em série
DNA	*Desoxyribonucleic Acid* - ácido desoxirribonucleico
PCR	*Polymerase Chain Rea*ction" - reação em cadeia da polimerase
LR	*Likelihood Ratio* - Razão de Verossimilhança
W	Probabilidade de Paternidade
HWE	Equilíbrio de *Hardy-Weinberg*
AMOVA	Análise da variância molecular
FBI	*Federal Bureau Investigation* - Agência Federal de Investigação
NDIS	*National DNA Index System* - Sistema de Índice Nacional de DNA
CODIS	*Combined DNA Index System* - sistema de índice de DNA combinado
IBDFAM	Instituto Brasileiro de Direito de Família
EDTA	Ácido etilenodiamino tetra-acético
BA	Bahia
ES	Espirito Santo
MG	Minas Gerais
USA	United States of America
PE	*Power of exclusion* - Poder de Exclusão
RMP	*Random match probability* - Probabilidade de Correspondência
PIC	*Polymorphism Information Content* - Informação Polimórfica Contida
bp	*Base Pair* - pares de bases
N	*North* – Norte
NE	*Northeast* – Nordeste
MD	*Midwest* – Centro-Oeste

S	*South* – Sul
SE	*Southeast* – Sudeste
SNP	*Single Nucleotide Polymorphism* – Polimorfismo de base única
PD	*Power of Discrimination* – Poder de discriminação
TPI	*Typical paternity index* – Índice Típico de Paternidade

RESUMO

Os Microssatélites ou *Short Tandem Repeats* (STR) vêm sendo amplamente usados em testes de paternidade e ciência forense desde meados dos anos 90. Os marcadores STRs altamente polimórficos são utilizados pela sua capacidade de diferenciar indivíduos. O teste de Paternidade é baseado em teste de hipóteses, onde o problema a ser resolvido é determinar se o suposto pai testado é realmente o verdadeiro pai da criança. Para a aplicação em genética forense além da padronização de novos marcadores é necessário também estudar a população na qual o teste será realizado. O Brasil tem uma população tri-híbrida, caracterizada por uma contribuição europeia, africana e ameríndia. Investigar a dinâmica dos alelos de cada marcador nas populações naturais e elucidar a diversidade genética nas mesmas é crucial para entender a história evolutiva e aplicar em estudos forenses. No presente trabalho, buscamos traçar o perfil genético da população brasileira utilizando um novo painel de dezoito marcadores utilizados em aplicações forenses. Para atingir nossos objetivos, caracterizamos molecularmente nove dos dezoito marcadores STR, avaliamos a informatividade dos dezoito marcadores através do cálculo de parâmetros forenses, definimos o perfil da população brasileira através de estudos da variabilidade genética e estimamos os níveis de contribuição europeia e africana nas populações por meio de estatísticas bayesianas. Todos os dados analisados neste trabalho foram obtidos de resultados de testes de verificação de parentesco realizados no laboratório Biocod Biotecnologia no triênio 2007/2008/2009. No capítulo I – "*Molecular characterization and population genetics of non-CODIS microsatellites used for forensic applications in Brazilian populations"* foi possível determinar o motivo de repetição de seis dos nove novos marcadores caracterizados através do sequenciamento dos alelos mais frequentes, os valores observados para os parâmetros de aplicação forense demonstram que o conjunto de marcadores estudados é tão informativo para elucidação de casos forenses quanto os marcadores do CODIS, os marcadores caracterizados apresentam baixa taxa de mutação e são uteis para diferenciar populações geneticamente. No capítulo II – "*Genetic profile and admixture of the Brazilian population based in markers used for forensic applications*" as populações brasileiras foram divididas de acordo com as regiões geográficas (Norte, Nordeste, Centro-Oeste, Sudeste e Sul) e foram comparadas com amostras parentais do painel SNP500 Câncer do repositório Coriell (Africana, Europeia e Hispanica); os resultados demonstraram que os marcadores previamente caracterizados são informativos tanto para análises forenses quanto para estudo genético-populacionais e que as populações Brasileiras receberam uma maior contribuição europeia do que africana e são geneticamente diferentes. Recentemente, outros conjuntos de STRs autossômicos vêm ganhando destaque para aumentar as chances de resolução de casos complexos, casos onde o suposto é falecido ou está ausente, de verificação de parentesco. Ao final podemos concluir que os marcadores caracterizados são bons marcadores para elucidar casos forenses por se mostraram tão informativos quanto os marcadores do sistema CODIS e outros marcadores previamente validados. As análises destes novos marcadores auxiliarão na resolução de casos complexos de verificação de parentesco e casos post-mortem. Nas análises populacionais foi possível verificar diferenças genéticas significativas entre as populações brasileiras. Ainda nas análises populacionais foi possível confirmar que a contribuição genética europeia foi maior que a africana durante o processo de formação da população brasileira.

ABSTRACT

Microsatellite or Short Tandem Repeats (STR) has been widely used in paternity testing and forensic science since the mid-90. Highly polymorphic STR markers are used for their ability to differentiate individuals. The Paternity test is based on hypothesis testing, where the problem to be solved is to determine whether the alleged father is the real father of the child. For forensic genetics, besides the standardization of new markers, it is also necessary to study the population in which the test will be performed. Brazil has a three-hybrid population, characterized by a European, African and Amerindian ancestry components. Understanding the molecular basis of allelic diversity for STR may be helpful to understand the evolutionary history of populations and for forensic applications. In the present work, we trace the genetic profile of the population using a new panel of 18 markers . To achieve our goal, we characterize molecularly nine of the eighteen STR markers; we evaluate the informativeness of eighteen markers by calculating forensic parameters, define the profile of the Brazilian population through studies of genetic variability and estimate the European and African contribution levels in populations through Bayesian statistics. All data analyzed in this study were obtained from results of kinship verification tests in Biocod Biotechnology laboratory in the three years 2007/2008/2009. In Chapter I - "Molecular characterization and population genetics of non-CODIS microsatellites used for forensic applications in Brazilian Populations" it was possible to determine the repeating motif of six of the nine markers by sequencing of the most frequent alleles, the observed values for the forensic parameters show that the set of studied markers are informative for elucidation of criminal cases as the CODIS markers. The markers characterized have low mutation rate and are useful to differentiate populations genetically. In Chapter II - "Genetic profile and admixture of the Brazilian population based in markers used for forensic applications", Brazilian populations were divided according to geographical regions (North, Northeast, Midwest, Southeast and South) and were compared with samples the parental SNP500 Cancer Coriell repository (African, European and Hispanic) panel; the results showed that the markers previously characterized are informative for both forensic analysis as to population genetic studies. Brazilian populations received more European than African contribution and are slightly different between the regions of the country. We conclude that characterized markers are good markers to elucidate forensic cases since proved as informative as markers of CODIS system and other markers validated in other studies and that these markers are useful for population genetic studies.

1 – INTRODUÇÃO

1.1 – Microssatélites ou *Short Tandem Repea*ts (STR)

Os Microssatélites ou *Short Tandem Repeats* (Repetições curtas em série) vêm sendo amplamente utilizados em testes de paternidade e ciência forense desde meados dos anos 90 (Weir et al., 2006). Estão entre os marcadores de DNA mais polimórficos do genoma e podem ser classificados de acordo com número de nucleotídeos no motivo de repetição di-, tri-, tetra-, penta- ou hexanuclutotídeos (Ellegren, 2004). Os STRs usados como marcadores genéticos para identificação individual estão em regiões de DNA não-codificantes e que seguem o modelo de evolução neutra. A variação genética nos locos STRs é caracterizada pela alta heterozigosidade e a presença de múltiplos alelos (Ellegren, 2004). Além disso, permitem a genotipagem num curto período de tempo e, ainda, são eficazes na identificação de amostras degradadas (Cabrero *et al.*, 1995).

A taxa de mutação dos STRs, de uma a duas mutações a cada 1.000 gerações, é devido à arquitetura molecular destes marcadores. Os STRs seguem, normalmente, um tipo específico de mutação - *step-wise mutation model*: adição ou subtração de uma unidade repetitiva. Esse processo acontece durante a replicação de uma nova fita de DNA, a polimerase desassocia-se transitoriamente da fita molde e volta a se associar de maneira errada (Sun *et al.*, 2014; Ellegren, 2004). Para os STRs, também é observado que a taxa de mutação pode aumentar com o aumento no tamanho do alelo, sendo comum observar mutações de duas ou mais sequências repetitivas (Balding, 2005).

Outra característica dos STRs é o fato de permitirem a amplificação simultânea por PCR em multiplex. Para a amplificação em multiplex é necessário agrupar os STRs de acordo com o tamanho do produto de PCR e suas diferentes marcações fluorescentes. Neste tipo de análise se consegue um alto poder de discriminação sem consumo de grande quantidade de DNA (Butler, 2007). A análise de vários marcadores moleculares aumenta a confiabilidade nas inferências dos casos de análises de parentesco e resolução de crimes.

Os sistemas multiplex são analisados em plataformas automatizadas de equipamentos de sequenciamento, baseados na eletroforese capilar com múltiplos canais usados para detectar produtos de PCR marcados com diferentes fluorescências (Jobling & Gill, 2004).

A resolução de testes de verificação de parentesco e casos forenses é composta pela genotipagem dos STRs corroborada por uma interpretação estatística dos resultados. Os marcadores STRs altamente polimórficos são utilizados pela sua capacidade de diferenciar indivíduos. Para se determinar as frequências alélicas destes marcadores são realizados estudos populacionais, com populações de diferentes grupos ancestrais e regiões geográficas (Huston, 1998).

1.2 - Teste de Paternidade e Verificação de Parentesco

O teste de paternidade se baseia em princípios básicos da genética: leis de Mendel e alta variabilidade genética. Cada indivíduo possui dois alelos para cada loco e os pares diferentes se distribuem independentemente na formação dos gametas. Na formação do zigoto metade da informação genética do indivíduo é herdada de sua mãe e a outra metade herdada de seu pai. O teste de paternidade consiste em uma comparação entre os alelos encontrados nos filhos e nos supostos pais, onde a presença de alelos paternos no material genético do filho é o primeiro indício de paternidade.

O teste de Paternidade é baseado em teste de hipóteses, onde o problema a ser resolvido é determinar se o suposto pai testado é realmente o verdadeiro pai da criança. Para resolver este problema é necessário calcular a razão de verossimilhança (*Likelihood Ratio* – LR) entre duas hipóteses testadas H_0 e H_1 (Gjertson *et al.*, 2007).

H_0: O suposto pai é o pai da criança

H_1: O suposto pai não é o pai da criança

$$LR = H_0/ H_1$$

Sendo assim, podemos citar como exemplo um caso de trio (mãe, filho e suposto pais) onde temos o perfil genético de cada um para o marcador D3S1358:

A mãe possui o genótipo 13/15, o suposto pai 16/17 e o filho 13/16, nesse caso, observa-se que o alelo 13 foi herdado da mãe e o alelo 16 do pai, e que o pai testado possui o alelo procurado. Então, a partir dessa informação podemos calcular a razão de verossimilhança entre as hipóteses:

$$H_0: 2xf(13)xf(15) \text{ x } ½ \text{ x } 2xf(16)xf(17) \text{ x } ½$$

$$H_1: 2xf(13)xf(15) \text{ x } ½ \text{ x } 2xf(16)xf(17)\ xf(16)$$

$$LR: ½ / f(16)$$

Nesse caso para H_0, hipótese do suposto pai testado ser o pai da criança, calcula-se a probabilidade do genótipo da mãe, a probabilidade do genótipo do suposto pai e a probabilidade da mãe ter passado um alelo para o filho e a probabilidade do pai ter passado o outro alelo para o filho. Para H_1, hipótese do suposto pai ser qualquer homem na população, calcula-se a probabilidade do genótipo da mãe, a probabilidade do genótipo do suposto pai e a probabilidade da mãe ter passado um alelo para o filho e a probabilidade do outro alelo do filho ter sido herdado de qualquer outro homem aleatoriamente na população. A razão de verossimilhança (LR) é calculada a partir da divisão H_0 por H_1 e demonstra quantas vezes é mais provável que o suposto pai em questão seja pai da criança.

A informação de um só marcador genético não é suficiente para se concluir sobre a probabilidade de paternidade, com isso a análise de vários marcadores em multiplex permite aumentar a informação de cada caso e assim calcula-se o LR combinado de todos os marcadores testados e com isso chega-se a uma Probabilidade de Paternidade (W) (Gjertson *et al.*, 2007).

$$W = LR/(1+LR)$$

1.3 - População Brasileira

O Brasil tem uma população tri-híbrida, caracterizada por uma contribuição europeia, africana e ameríndia. No início do século XVI estimava-se que mais de dois milhões de indígenas povoavam o Brasil, esse número foi reduzido devido às batalhas com os colonizadores e às doenças transmitidas por eles. No final do século XX o número de habitantes indígenas chegava a 302.888. A colonização portuguesa iniciou-se em 1500, mas o fluxo realmente aumentou nos períodos de 1760-1791 e de 1837-1841, cerca de 10 mil imigrantes. Os escravos negros chegaram ao Brasil a partir de 1701 originados da África Centro-Ocidental (hoje região ocupada por Angola). A partir de 1800 a grande maioria dos cinco milhões de imigrantes que chegaram ao Brasil era de origem portuguesa e italiana, seguidos por espanhóis, alemães, sírio-libaneses e japoneses (IBGE).

No contexto genético, este legado da história contribui para o aumento da heterogeneidade e um desbalanço nas frequências alélicas e genotípicas entre a população resultante e as principais populações fundadoras. Os níveis de ancestralidade genômica na população brasileira atual têm sido investigados extensamente em pesquisas que envolvem marcadores moleculares de diversas classes. Estudos mostram que a população brasileira é geneticamente heterogênea, porém com predominância europeia em seus marcadores autossômicos e, ainda corroboram com dados históricos com a observação de linhagem patriarcal tipicamente Europeia e matriarcal tri-parental, com grande influência de indígenas e africanos (Lins, 2007).

1.4 - Diferenciação Genética entre Populações

A Genética de Populações visa à investigação da dinâmica dos alelos nas populações naturais buscando a elucidação dos mecanismos que alteram a sua composição gênica (efeito de fatores evolutivos, isto é, mutação, seleção natural, deriva genética e fluxo gênico de populações migrantes) ou a frequência genotípica pelo aumento da homozigose (efeito dos casamentos consanguíneos ou da subdivisão da população).

Elucidar a diversidade genética nas populações humanas é crucial para entender sua história evolutiva (Scliar *et al.*, 2012). Estudos indicam que 5-10% da diversidade genética humana é explicada por diferenças genéticas entre as grandes regiões geográficas. Estes resultados indicam que existem mais similaridades do que diferenças entre populações humanas geograficamente distintas (Holsinger & Weir, 2009).

Populações naturais, incluindo as populações humanas, possuem geografia e história complexas. Estudar como as populações são formadas é difícil e a abordagem mais tradicional destas análises é fundamentalmente por modelos matemáticos que determinam a estrutura das populações (Hey & Machado, 2003).

O Equilíbrio de *Hardy-Weinberg* (HWE) é princípio matemático clássico em genética de populações que descreve as frequências esperadas de genótipos para um loco após uma geração de cruzamentos casuais, a partir das frequências alélicas na população. O equilíbrio pode não se manter em populações reais, mas ele pode apresentar boas aproximações se o tamanho populacional for grande, se os casamentos forem ao acaso, e se não houver uma sobrevivência diferencial de zigotos com um genótipo específico para um determinado loco (Balding, 2005). Se compararmos as frequências genotípicas de uma população real com relações de Hardy-Weinberg, caso elas se desviem, isso sugere que eventos tais como seleção ou ausência de cruzamentos aleatórios possa agir sobre estas populações (Ridley, 2006).

A AMOVA (Análise da Variância Molecular) foi inicialmente introduzida como extensão às análises das frequências alélicas e reflete a correlação entre a diversidade entre diferentes níveis de subdivisão populacional. Essas análises fornecem informações sobre a estrutura genética das populações (Michalakis & Excoffier, 1996) determinada pela soma dos fatores que governam as forças pelas quais os gametas se unem para formar os zigotos da próxima geração. Uma das formas de se medir esta variância é através das estatísticas F descritos por Wright (Wright, 1951; Excoffier *et al.*, 1992; Bossart & Prowell, 1998).

Wright (1951) introduziu três parâmetros inter-relacionados para descrever a estrutura genética de populações. Estes parâmetros são: F_{it}, a correlação entre gametas dentro de um indivíduo relativo a toda a população; F_{is}, a correlação entre gametas dentro de um indivíduo relativo à subpopulação a qual esse indivíduo pertence; e F_{st}, a correlação entre gametas escolhidos randomicamente em uma mesma subpopulação relativa à totalidade da população ou como a proporção da diversidade genética devido a diferenças de frequência alélicas entre as populações (Holsinger & Weir, 2009).

Além de se determinar a estrutura genética das populações é possível também classificar indivíduos com origem genética desconhecida como pertencentes às populações previamente definidas. A definição de populações é tipicamente subjetiva, podendo ser definida de acordo com padrões linguísticos, culturais ou físicos, assim como a localização geográfica dos indivíduos amostrados. Após estimar as frequências alélicas das populações definidas calcula-se a probabilidade de um dado genótipo ser originado em cada população. Indivíduos de origem desconhecida podem ser atribuídos às populações de acordo com estas probabilidades (Pritchard *et al*, 2000b).

Segundo Pritchard (2000b), o modelo utilizado no programa STRUCTURE foi baseado em métodos de agrupamentos de dados de genotipagem multiloco para inferir a estrutura das populações e atribuir indivíduos a essas populações. Nesse modelo existem k populações (onde k pode ser desconhecido), cada uma delas é caracterizada por um conjunto de frequências alélicas para cada loco. Indivíduos de uma mesma amostra são atribuídos para uma população, ou reunidos em duas ou mais populações se seus genótipos indicarem que são miscigenados. Este modelo não assume um processo particular de mutação, e por isso pode ser aplicado para a maioria dos marcadores genéticos utilizados comumente, desde que eles não estejam ligados (marcadores localizados em regiões cromossômicas próximas que não são separadas durante o processo recombinação). Entre as aplicações desse modelo inclui-se: demonstrar a presença de estruturação nas populações, atribuir indivíduos a uma determinada população, estudo de zonas híbridas e identificar migrações e miscigenação.

2 – HIPÓTESE E JUSTIFICATIVA

A Genética Forense é a área que trata da utilização dos conhecimentos e das técnicas de genética e de biologia molecular no auxílio à justiça. O ramo mais desenvolvido da Genética Forense é a Identificação Humana pelo DNA e sua aplicação mais popular é o teste de paternidade. A evolução da genética forense foi impulsionada pela análise da variação genética humana, iniciou-se há mais de um século com a descoberta do polimorfismo dos grupos sanguíneos ABO por Karl Landsteiner e a percepção de que essa era uma ferramenta para elucidação de casos criminais (Jobbing & Gill, 2004).

A revolução do DNA iniciou-se em 1984 com a descoberta, por Alec Jeffreys, das regiões hipervariáveis conhecidas como minissatélites. Estes são detectados através da técnica de hibridização por sondas *Southern Blot*, que ficou conhecida como impressão digital do DNA. Essa técnica foi utilizada para resolver os primeiros casos criminais pela análise do DNA (Jobling & Gill, 2004).

A partir de 1988 a descoberta da técnica de PCR por Mullis & Faloona, proporcionou um aumento na sensibilidade, permitindo a amplificação de DNA degradado e a partir de então se tornou a base para as análises forenses. Em 1991 foi descoberto o primeiro STR, marcador multi-alélico e com padrão de herança codominante (Jobling & Gill, 2004). As vantagens obtidas após as duas descobertas abriram caminho para a criação de bancos de dados nacionais.

O laboratório do FBI, nos Estados Unidos da América, foi o pioneiro na criação deste tipo de banco de dados com o desenvolvimento do sistema combinado de índices de DNA (CODIS), que combina a Ciência Forense e a Tecnologia da Informática, proporcionando uma ferramenta efetiva para o desenvolvimento da investigação criminal. O sistema CODIS é composto por 13 locos: CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, TH01, TPOX e vWA. Tal sistema permite a todos os laboratórios americanos (federais, estaduais e locais) realizarem permutas e comparações dos perfis de DNA eletronicamente, além de possibilitar a interligação dos crimes entre si e dos suspeitos envolvidos (FBI).

A Genética forense teve grande avanço nos últimos 20 anos, após o início da utilização dos marcadores STRs. A facilidade de análise destes marcadores permitiu uma melhoria nos serviços de identificação humana e pesquisas na área. A divulgação dos marcadores do sistema CODIS facilitou a realização dos testes de verificação de parentesco, já que estes marcadores são analisados em todos os testes deste tipo, e abriu caminho para uma enorme quantidade de publicações com estes STRs em diferentes populações, incluindo a população brasileira (Sousa *et al.*, 2014; Aguiar *et al.*, 2012).

A genética forense no Brasil vem sendo impulsionada por dois fatores: a- em 2009, esforços visando o desenvolvimento da Genética Forense no cenário nacional resultaram na assinatura do Termo de Compromisso para utilização do software CODIS no Brasil (Aguiar *et al*., 2011); e b- de acordo com o projeto de lei apresentado para regulamentação do exame de DNA no Brasil, 30% dos registros de nascimento feitos no Brasil não tem o nome do pai, o que corresponde a quase um milhão de nascimentos por ano e implica em um crescente número de ações de investigação de paternidade/maternidade (IBDFAM). Com o aumento de ações de investigação de paternidade, a procura pelos testes de verificação de parentesco também aumentou bastante. A maior parte dos casos é resolvida pelos testes simples, TRIO (mãe, filho e suposto pai) e DUO (filho e suposto pai), porém testes complexos começaram a ser solicitados para resolver casos onde o suposto está ausente ou faleceu. Para a resolução destes casos complexos, muitas vezes o número de marcadores disponibilizados em kits comerciais não é suficiente, faz-se necessário a utilização de marcadores adicionais. As análises de marcadores adicionais requerem estudos preliminares que caracterizem estes marcadores e determinem o perfil genético da população onde o teste será realizado.

No presente trabalho, verificamos se os novos marcadores selecionados são tão informativos para análises forenses e estudos populacionais que os marcadores do sistema CODIS, para isso buscamos traçar o perfil genético da população brasileira utilizando um painel de dezoito marcadores utilizados em aplicações forenses. Para atingir nossos objetivos, caracterizamos molecularmente os nove dos dezoito marcadores STRs que ainda não foram caracterizados, avaliamos a informatividade dos dezoito marcadores STRs através do cálculo de parâmetros forenses, definimos o perfil da população brasileira através de estudos da variabilidade genética e estimamos os níveis de contribuição europeia e africana na população brasileira por meio de estatísticas bayesianas.

Todos os dados analisados neste trabalho foram obtidos de resultados de testes de verificação de parentesco realizados no laboratório Biocod Biotecnologia no triênio 2007/2008/2009.

2.1 – Biocod Biotecnologia

A Biocod Biotecnologia é um laboratório especializado em análises genéticas com mais de 10 anos de experiência. Para a realização destes testes, a Biocod conta com uma equipe técnica especializada e com mais de 15 anos de experiência, além de equipamentos de última geração.

O teste mais difundido, dentre os vários disponíveis na Biocod, é o teste de paternidade. Mensalmente são realizados aproximadamente 700 casos de investigação de vínculo genético. Devido ao grande número de testes de paternidade, a Biocod conta com um banco de dados genéticos com mais de 40.000 indivíduos, o que possibilita inúmeras análises genéticas.

2.2 – Teste de Paternidade na Biocod

Assim que chegam ao laboratório, todas as amostras são inspecionadas, codificadas e cadastradas no banco de dados. Para cada amostra são cadastrados os seguintes dados: Nome, Endereço, Data e Local de Nascimento e de Coleta, Sexo e Tipo de Exame. São recebidos diariamente três tipos de amostras: sangue coletado em tubos com EDTA, sangue coletado em papel de filtro tipo *FTA card* (Whatman®) e esfregaço de células da mucosa bucal conservadas em álcool. Em alguns casos mais raros, podem ser recebidas também amostras de vilo corial, biópsia de tecidos em geral e material exumado.

Para cada tipo de amostra é seguido um protocolo de extração diferente, visando a uma extração de DNA rápida, em concentrações suficientes para análises de qualidade e com um custo reduzido. A extração de sangue coletado em tubo com EDTA e células da mucosa bucal é feita com base em protocolos *salting out*, podendo raramente seguir protocolos que utilizem o fenol-clorofórmio. A extração de sangue em *FTA card* (Whatman®) é simples e baseia-se na lavagem das impurezas do papel deixando o DNA impregnado no mesmo.

Após a obtenção de DNA de qualidade, as amostras são amplificadas através da técnica de PCR-multiplex, onde várias regiões do DNA são amplificadas em uma mesma reação, reduzindo tempo e custo das análises.

Rotineiramente são amplificados dois destes painéis, PAINEL 1 e PAINEL 2, que contam com 18 marcadores STRs (tabela 1), mesclando marcadores do sistema CODIS, marcadores caracterizados em estudos prévios (Wenda *et al.,* 2005; Garofano *et al*, 1999, Lareu *et al*., 1996) e marcadores ainda não utilizados para este fim.

STRs	**Painel**	**Informação Molecular**	**Tamanho**
**D2S1338	PAINEL 1	Perfeito	165-205
*D3S1358	PAINEL 1	Perfeito	123-143
D3S2387	PAINEL 1	Composto	177-209
D3S2406	PAINEL 1	Composto	306-350
D5S2503	PAINEL 1	Perfeito	350-390
*D5S818	PAINEL 1	Imperfeito	120-150
*D7S820	PAINEL 2	Perfeito	204-240
D9S938	PAINEL 1	Perfeito	369-421
D10S1237	PAINEL 2	Perfeito	376-432
**D12S391	PAINEL 1	Imperfeito	211-251
*D13S317	PAINEL 2	Perfeito	175-199
*D16S539	PAINEL 2	Perfeito	148-172
D16S753	PAINEL 1	Composto	252-276
D21S1437	PAINEL 2	Perfeito	119-143
D22S534	PAINEL 1	Perfeito	450-515
D22S689	PAINEL 1 E 2	Composto	202-226
**SE33	PAINEL 2	Composto	197-343
*TH01	PAINEL 2	Imperfeito	146-190

Tabela 1 – Caracterização dos Marcadores STRs presentes nos dois painéis de acordo com a informação molecular e o tamanho do produto de PCR. *Marcadores pertencentes ao sistema CODIS; **Marcadores caracterizados em estudos prévios.

As amostras amplificadas são genotipadas por eletroforese capilar em sequenciador *MegaBACE* 1000 (GE Healthcare) e são analisadas pelo software *Fragment Profile* v2.0 (GE Healthcare).

O envio de dados genotípicos para o banco de dados no módulo do Laboratório de Paternidade acontece no momento da liberação dos resultados. Para essa liberação é realizada uma conferência dos resultados das genotipagens. Após essa conferência, os marcadores que não apresentaram bons resultados são retirados da análise e os seus perfis genéticos não são enviados ao banco de dados, o que justifica um número diferente de indivíduos para cada marcador.

2.3 – Banco de dados

Foram selecionados indivíduos não aparentados envolvidos em casos de TRIO e DUO. Para eliminar a consanguinidade, nenhum dos filhos foi considerado neste estudo. De cada indivíduo foram extraídas as seguintes informações: Indivíduo, Cidade Naturalidade, Sigla Estado Naturalidade, Tipo de Coleta, Local de Coleta, Tipo de Contrato e Genótipo para os marcadores escolhidos.

Os dados foram divididos em quatro conjuntos diferentes (Tabela 2): Dados 1 - todos os indivíduos não aparentados e com no mínimo 15 marcadores genotipados (D10S1237, D12S391, D13S317, D16S753, D21S1437, D22S534, D2S1338, D3S1358, D3S2387, D3S2406, D5S2503, D7S820, D9S938, SE33 e TH01); Dados 2 - todos os indivíduos não aparentados e com no mínimo 12 marcadores genotipados dentre os 15 marcadores mais frequentes no banco de dados; Dados 3 - todos os indivíduos não aparentados, com no mínimo 15 marcadores genotipados e com informação de cidade e estado naturalidade; e Dados 4 - Todos os indivíduos não aparentados, com no mínimo 12 marcadores genotipados dentre os 15 frequentes no banco de dados e com informação de cidade e estado naturalidade.

	Dados 1	**Dados 2**	**Dados 3**	**Dados 4**
Total	11.241	21.802	3.251	7.095
Número mínimo de Marcadores	15 marcadores	12 marcadores dos 15 mais comuns	15 marcadores	12 marcadores dos 15 mais comuns
Características	-	-	Cidade Naturalidade	Cidade Naturalidade

Tabela 2 – Distribuição do número de indivíduos em cada Banco de Dados pré-definidos.

2 - OBJETIVOS:

2.1 - GERAL

- Traçar o perfil genético da população brasileira a partir de um novo painel de dezoito marcadores STR utilizados em aplicações forenses**.**

2.2 - ESPECÍFICOS

- Determinar a estrutura molecular dos marcadores: D3S2387, D3S2406, D5S2503, D9S938, D10S1237, D16S753, D21S1437, D22S534 e D22S689.

- Avaliar os parâmetros forenses para os dezoito marcadores do painel da Biocod Biotecnologia: frequência alélica, poder de exclusão, probabilidade de correspondência, poder de discriminação, conteúdo de informação do polimorfismo e índice típico de paternidade.

- Calcular a taxa de mutação de cada marcador do painel da Biocod Biotecnologia.

- Verificar a variabilidade genética da população brasileira, após uma subdivisão de acordo com as regiões geográficas, através da análise da variância molecular.

- Estimar a contribuição de populações de origem europeia e africana na população brasileira.

3 - RESULTADOS

3.1 - Capítulo I – *Molecular characterization and population genetics of non-CODIS microsatellites used for forensic applications in Brazilian populations*.

PINTO, LAÉLIA MARIA, OLIVEIRA, CRISTIANE LOMMEZ DE, SANTOS, LUCIANA LARA DOS, TARAZONA-SANTOS, EDUARDO Molecular characterization and population genetics of non-CODIS microsatellites used for forensic applications in Brazilian populations. *Forensic Science International: Genetics* 9 (2014) e16-e17.

A caracterização de novos marcadores é importante para perícias nas quais apenas os marcadores CODIS não são suficientes para a finalização dos casos. Este estudo teve como principais objetivos: i) caracterizar molecularmente os STRs D3S2387, D3S2406, D5S2503, D9S938, D10S1237, D16S753, D21S1437, D22S534 e D22S689; ii) calcular os parâmetros estatísticos que demostram a informatividade de cada um dos dezoito marcadores: poder de exclusão, probabilidade de coincidência, informação polimórfica contida no marcador, taxas de mutação e as frequências alélicas; iii) verificar através do painel da Biocod Biotecnologia a variabilidade genética humana e a diferenciação genética entre as subpopulações. Para caracterização molecular foram sequenciados indivíduos homozigotos para os dois alelos com maior frequência na população brasileira. As amostras selecionadas para definir a informatividade dos marcadores e demostrar a variabilidade genética foram extraídas do Banco de Dados 4 da Biocod Biotecnologia. A diferenciação genética entre subpopulações foi realizados apenas com as subpopulações do três estados brasileiros com maior número de indivíduos disponíveis: Bahia, Espirito Santo e Minas Gerais. A partir dos resultados dos sequenciamentos foi possível determinar o motivo de repetição de seis dos noves marcadores caracterizados (*Supplementary Table 1*). Os valores observados para os parâmetros de aplicação forense demonstram que o conjunto de marcadores estudados é tão informativo para elucidação de casos de paternidade, identificação humana e casos post-mortem quanto os marcadores do CODIS (*Supplementary Figure 1* e *Supplementary Table 3*). Os marcadores caracterizados neste estudo apresentam baixa taxa de mutação (*Supplementary Table 5*). Os resultados dos cálculos estatísticos sugere que estes marcadores podem ser usados para análises em diferentes populações (*Supplementary Table 2* e *Supplementary Table 4*).

MOLECULAR CHARACTERIZATION AND POPULATION GENETICS OF NON-CODIS MICROSATELLITES USED FOR FORENSIC APPLICATIONS IN BRAZILIAN POPULATIONS

Laélia Maria Pinto[1,2], Cristiane Lommez de Oliveira[1], Luciana Lara dos Santos[3], Eduardo Tarazona-Santos[2]

[1]Biocod Biotecnologia. Avenida do Contorno 9636 3th floor, Santo Agostinho, Belo Horizonte, MG, zip code 30110936, Brazil.

[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Post office box 486, Belo Horizonte, MG, zip code 31270-910, Brazil.

[3]Universidade Federal de São João Del Rei – Campus Centro-oeste Dona Lindu. Rua Sebastião Gonçalves Coelho, 400, Divinópolis, MG, zip code 35501-296, Brazil.

CORRESPONDING AUTHOR:

Laélia Maria Pinto

Rua São Paulo 47

Conjunto São José, Esplanada, Sabará, MG, zip code 34515320, Brazil

Phone: 55 31 36726993, Fax: 55 31 30365002

Email: laeliap@hotmail.com

Dear Editor,

Several microsatellites PCR (polymerase chain reaction) multiplex systems (i.e. for simultaneous typing) have been reported for forensic analysis. These include autosomal STR multiplex kits widely used which are commercially available. These commercial kits are generally based on the Combined DNA Index System (CODIS) loco, and a huge volume of genetic population data for the CODIS loco from different ethnic groups has been reported [1-4]. Nevertheless, there are hundreds of other highly polymorphic STR loco unlinked to the current CODIS loco (non-CODIS markers), which are also useful for forensic genetics [5]. Analysis of further non-CODIS STR loco may complement information from CODIS locus, offering powerful tools for difficult kinship testing, such as sib-ship testing or testing for deficient paternity cases [6].

The Brazilian population is characterized by high ethnic variability produced by admixture, which turned it suitable for genetic diversity studies from a forensic perspective. The Biocod STR Database was constructed based on individuals from all Brazilian States who participated in paternity tests (realized in Biocod Biotecnologia). In this study we analyze data from three States for which large number of individuals are available (a total of 5,639 individuals): Bahia (BA, Northeast Brazil, with a high level of African admixture, n= 2,891), Espírito Santo (ES, n= 1,534) and Minas Gerais (MG, n= 1,214), both from the South East of the country and with intermediate level of admixture [7]. We analyzed two STR-multiplex (Panel 1 and Panel 2) with a total of 18 markers used in the Biocod lab routine. These markers are divided in two groups: CODIS markers - D3S1358 (Panel 1), D5S818 (Panel 2), D7S820 (Panel 1), D13S317 (Panel 1), D16S539 (Panel 1) and TH01 (Panel 1); and non-CODIS markers - D2S1338 (Panel 2), D3S2387 (Panel 2), D3S2406 (Panel 2), D5S2503 (Panel 2), D9S938 (Panel 1), D10S1237 (Panel 1), D12S391 (Panel 2), D16S753 (Panel 2), D21S1437 (Panel 1), D22S534 (Panel 2), D22S689 (Panels 1 and 2) and SE33 (Panel 1). The research performed follows the ISFG guideline [8].

By resequencing of homozygous individuals, we analyzed for the first time six non-CODIS markers, determining their repeat counts and its sequence: D3S2387, D3S2406, D9S938, D10S1237, D22S534 and D22S689. We determined that all markers of our study are tetranucleotides (Supplementary Table 1). Moreover, four out of six re-sequenced markers have a simple structure, containing uninterrupted runs of units sharing a homogenous array. Two markers are complex, with an interruption or change in the sequence. We did not obtain the sequence of 3 non-CODIS markers (D5S2503, D16S753 and D21S1437), likely because the size of amplicons were too small (100-200pb) or the primer design was not appropriate for sequencing.

The population genetics analysis performed using the software Arlequin [9] showed that all population were in Hardy-Weinberg equilibrium for all loco after performing the Bonferroni correction (P was always higher than 0.002). Consistently with the history of recent admixture of the studied populations, we observed linkage disequilibrium between some of the markers (Supplementary Table 2)[10-12].

By F_{ST} analysis [13] we observed that the populations from Minas Gerais and Espirito Santo, both in South Eastern Brazil, are not differentiated (F_{ST}= 0.00002, P=0.32), while Bahia is significantly differentiated both from Espirito Santo (F_{ST}=0.00028, P=0.00000) and Minas Gerais (F_{ST}=0.00062, P=0.00000) (Supplementary Table 3), consistently with its historically reported higher African ancestry, that has also been confirmed by several population genetic studies [7].

For each STR studied, CODIS and non-CODIS, we used the PowerStats v.1.2 [14] (Promega Corporation, Madison, WI, USA) software to estimate the Power of exclusion (PE), the Random match probability (RMP), the Polymorphism Information Content (PIC) and allelic frequencies (Supplementary Table 4). Overall, the averages for the three forensic statistics calculated over loco (Supplementary Figure 1) are similar in the populations of Bahia, Espírito Santo and Minas Gerais. For the three considered parameters non-CODIS markers of this study show values that are comparable to those observed for CODIS markers across the three studied populations (Supplementary Figure 1). Among the studied markers, SE33 (that has 56 alleles in our sample) is the most informative and D22S534 (that has 12 alleles in our sample) is the less informative.

We estimated for the two combined panels 1 and 2 a combined PE of 0.999967, a combined RMP of 4.036 x 10^{-24} and an average PIC of 0.795. STRs are considered informative if they have RMP values below 0.1. Most of the STR of our panel presented values below this, and only three markers (D16S539, D5S818 and D22S534) in each population presented values close to 0.1.

The high mutation rate in microsatellite loco allows mutation events to be directly observed, provided that an enough number of meiosis is evaluated [15]. We can observe that the markers of our study have a low mutation rate in comparison with other studies [16]. SE33 presents the highest mutation rate (0.0022) and the TH01 did not show mutations (Supplementary Table 5). Interestingly, markers with the higher mutation rates have a complex sequence.

Our set of 18 markers, routinely used for forensic analysis by the BIOCOD laboratory in analysis of paternity cases as well as in human identification and *post-mortem* cases, is at least as much informative as a CODIS panel of 13 STR, which is the minimal number of markers established by the forensic community to resolve forensic cases. The panels used by BIOCOD have been used to resolve more than 80,000 forensic analyses coming from all across the country. We suggest that these markers may be used in forensic analysis in different European and Latin American populations.

References

[1] E. Chouery, M.D. Coble, K.M. Strouss, J.L. Saunier, N. Jalkh, M. Medlej-Hashim, F. Ayoub, A. Mégarbané , Population genetic data for 17 STR markers from Lebanon, Legal Med 12 (2010) 324-326.

[2] L.N. Xu, S.P. Hu, G.Y. Feng, STR polymorphisms of the Henan population and investigation of central plains Han origin of Chaoshanese, Biochem Genet 47 (2009) 569-581.

[3] D. Grattapaglia, A.B. Schmidt, C. Costa e Silva, C. Stringher, A.P. Fernandes, M.E. Pereira., Brazilian population database for the 13 STR loco of the Ampf/STR® Profile Plus^{TM} and Cofiler^{TM} multiplex kits, Forensic Sci. Int 118 (2001) 91-94.

[4] B. Egyed, S. Füredi, M. Angyal, I. Balogh, L. Kalmar, Z. Padar, Analysis of the population heterogeneity in Hungary using fifteen forensically informative STR markers, Forensic Sci. Int 158 (2006) 244-249.

[5] H. Asamura, M. Ota, H. Fukushima, Population data on 10 non-CODIS STR loco in Japanese population using a newly developed multiplex PCR system, J of Forensic and Legal Med. 15 (2008) 519-523.

[6] H. Asamura, S. Fujimori, M. Ota, H. Fukushima, MiniSTR multiplex systems based on non-CODIS loco for analysis of degraded DNA samples, Forensic Sci. Int 173 (2007) 7-15.

[7] S. R. Giolo, J. M. P. Soler, S. C. Greenway, M. A. A. Almeida, M. de Andrade, J. G. Seidman, C. E. Seidman, J. E. Krieger, A. C. Pereira, Brazilian urban population genetic structure reveals a high degree of admixture, Eur J Hum Genet 20 (2012) 111-116.

[8] A. Carracedo, J.M. Butler, L. Gusmão, W. Parson, L. Roewer, P.M. Schneider, Publication of population data for forensic purposes, Forensic Sci. Int. Genet. 4 (2010) 145:147.

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

[10] R.C. Lewotim, K. Kojima, The evolutionary dynamics of complex polymorphisms, Evolution 14 (1960) 450-472.

[11] M. Slatkin, Linkage disequilibrium in growing and stable population, Genetics 137 (1994a) 331-336.

[12] M. Slatkin, L. Excoffier, Testing for linkage disequilibrium in genotypic data using EM algorithm, Heredity 76 (1996) 377-383.

[13] M. Slaktin, A measure of population subdivision based on microsatellite alleles frequencies, Genetics 139 (1995) 457-462.

[14] PowerStats. A computer program for the analysis of population statistics (1999). Free program distributed by the authors over the internet from http://www.promega.com/geneticidtool.

[15] H. Ellegren, Microsatellites: simple sequences with complex evolution, Nat. Rev. Genet. 5 (2004) 435-445.

[16] E.S. Andrade, A.V. Gomes, G. Raposo, L. Mauricio-da-Silva, R.S. Silva, Mutation rates at 14 STR loco in the population from Pernambuco Northeast Brazil, Forensic Sci. Int. Genet. 3 (2009) e141-e143.

Supplementary Table 1

Supplementary Table 1 - Molecular characterization and chromosomal location of the 9 STR loci

Microsatellites	Alelle size (bp)	Alelle number (Repeats)	Molecular information	Chromosomal location	Reference
D3S2387[a]	196	22	$(GATA)_{12}$ $(GACA)_{10}$	3p26.3	Characterized in this study
D3S2406[a]	316	32	$(GGAT)_7$ $(GGAC)_6$ $(GACA)_8$ $(GATA)_{11}$	3p12	Characterized in this study
D5S2503[b]	354-382	-	GATA	5p14	http://alfred.med.yale.edu
D9S938[a]	400	26	$(GGAA)_{26}$	9q31	Characterized in this study
D10S1237[a]	404	20	$(GATA)_{20}$	10q25	Characterized in this study
D16S753[b]	252-276	-	GGAA	16p11.1	http://alfred.med.yale.edu
D21S1437[b]	111-151	-	GGAA	21q11.2	http://alfred.med.yale.edu
D22S534[a]	485	13	$(TACA)_{13}$	22q13	Characterized in this study
D22S689[a]	214	11	$(GATA)_{11}$	22q12	Characterized in this study

[a]STR loci characterized by sequencing.

[b]STR loci not characterized by sequencing

Supplementary Figure 1

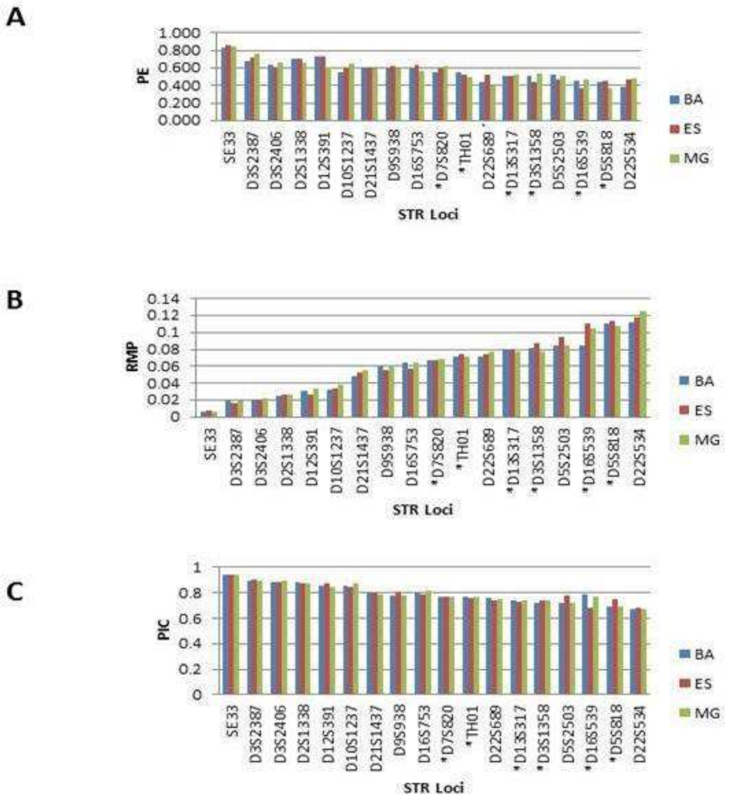


Fig. 1 - Forensic statistics: Power of Exclusion (PE), Random Match Probability (RMP) and Polymorphism Information Content (PIC), measures for 3 Brazilian population Bahia (BA), Espirito Santo (ES) and Minas Gerais (MG) for 18 STR loci. (A) PE, (B) RMP and © PIC. *CODIS markers.

Supplementary Table 2

Supplemntary Table 2 - Pairwise linkage disequelibrium by State, p-Values estimated from EM algorithm[10-12]. Population A: Bahia, B: Espirito Santo and C: Minas Gerais.

Population A: Bahia

Markers	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	TH01
D10S1237																		
D12S391	0.02346																	
D13S317	0.34800	0.17595																
D16S539	0.30596	0.90811	0.07918															
D16S753	0.00000	0.01173	0.04692	0.00000														
D21S1437	0.43793	0.11046	0.66960	0.01955	0.05767													
D22S534	0.27273	0.60313	0.68915	0.07331	0.02248	0.03812												
D22S689	0.00000	0.36266	0.14956	0.09873	0.00000	0.62561	0.00000											
D2S1338	0.16813	0.00000	0.17595	0.07234	0.06158	0.41740	0.09091	0.01857										
D3S1358	0.24731	0.49365	0.05083	0.00782	0.00000	0.00196	0.05181	0.00000	0.21310									
D3S2387	0.00000	0.03812	0.00196	0.00000	0.00000	0.10850	0.00000	0.00000	0.28446	0.00000								
D3S2406	0.00000	0.07625	0.20626	0.00000	0.00000	0.50440	0.20919	0.00000	0.00978	0.38514	0.00587							
D5S2503	0.32942	0.54545	0.44673	0.00196	0.00293	0.04790	0.00000	0.00000	0.82209	0.51417	0.02639	0.00000						
D5S818	0.00000	0.02151	0.07234	0.00000	0.00000	0.00000	0.00000	0.00000	0.14272	0.00000	0.00000	0.00000	0.00000					
D7S820	0.27077	0.56305	0.13490	0.18280	0.00000	0.03324	0.03226	0.01466	0.09482	0.17693	0.07038	0.21212	0.48289	0.03617				
D9S938	0.00000	0.06061	0.08895	0.00000	0.00098	0.01760	0.00098	0.00000	0.90420	0.00000	0.00000	0.00000	0.00000	0.00000	0.37634			
SE33	0.36755	0.06843	0.15445	0.28446	0.00000	0.37341	0.30010	0.00000	0.11926	0.01662	0.00098	0.06647	0.07234	0.00000	0.08016	0.53861		
TH01	0.28837	0.23460	0.07527	0.13001	0.25024	0.02151	0.46432	0.20430	0.02737	0.19746	0.18866	0.05963	0.93451	0.12610	0.16031	0.00489	0.14467	

Population B: Espirito Santo

Markers	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	TH01
D10S1237																		
D12S391	0.47214																	
D13S317	0.01369	0.04399																
D16S539	0.02835	0.54741	0.68035															
D16S753	0.00000	0.15347	0.00880	0.00000														
D21S1437	0.53177	0.92082	0.03715	0.21212	0.00782													
D22S534	0.38807	0.88368	0.35875	0.70968	0.00000	0.19941												
D22S689	0.00000	0.15445	0.19355	0.44282	0.00000	0.01857	0.22092											
D2S1338	0.00000	0.33431	0.43793	0.03910	0.67155	0.11730	0.18084	0.53275										
D3S1358	0.19159	0.33627	0.43402	0.00489	0.00000	0.05767	0.10557	0.00880	0.85533									
D3S2387	0.00000	0.05181	0.07722	0.00000	0.00000	0.00196	0.16227	0.00000	0.00587	0.00000								
D3S2406	0.00000	0.00587	0.16618	0.00000	0.00000	0.53470	0.00000	0.00000	0.08798	0.00978	0.00000							
D5S2503	0.60215	0.53568	0.77517	0.05865	0.00000	0.08309	0.06647	0.00000	0.02933	0.57771	0.00293	0.00000						
D5S818	0.00000	0.03128	0.00391	0.00000	0.00000	0.00000	0.00000	0.00000	0.00880	0.00000	0.00000	0.00000	0.00000					
D7S820	0.00391	0.47703	0.31281	0.08113	0.00000	0.28837	0.04399	0.00000	0.18964	0.13294	0.02933	0.86901	0.20723	0.00098				
D9S938	0.00000	0.39101	0.00098	0.00000	0.00000	0.62366	0.00782	0.00000	0.20235	0.00000	0.00000	0.00000	0.03910	0.00000	0.09580			
SE33	0.49756	0.04203	0.05572	0.00196	0.00391	0.03030	0.00293	0.17107	0.00489	0.79863	0.13978	0.49658	0.12023	0.00587	0.33627	0.00978		
TH01	0.16129	0.27957	0.03519	0.06843	0.39883	0.16129	0.10362	0.03910	0.10068	0.27468	0.02835	0.31867	0.49560	0.18377	0.57771	0.04399	0.34897	

Population C: Minas Gerais

Markers	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	TH01
D10S1237																		
D12S391	0.44379																	
D13S317	0.24633	0.48974																
D16S539	0.95210	0.22972	0.56207															
D16S753	0.00000	0.43988	0.15836	0.00000														
D21S1437	0.16618	0.43206	0.24829	0.31281	0.10850													
D22S534	0.89834	0.08602	0.89932	0.92473	0.09384	0.01075												
D22S689	0.03324	0.63245	0.38025	0.90811	0.00000	0.31769	0.01466											
D2S1338	0.71652	0.85826	0.26002	0.14467	0.09189	0.43695	0.41056	0.01075										
D3S1358	0.00098	0.58065	0.27077	0.12219	0.04692	0.59629	0.40665	0.56989	0.02542									
D3S2387	0.00000	0.13392	0.38710	0.00000	0.00000	0.52981	0.03030	0.00000	0.04594	0.00000								
D3S2406	0.09873	0.17791	0.42913	0.00000	0.00000	0.09580	0.04790	0.00000	0.80059	0.44184	0.04203							
D5S2503	0.12805	0.13490	0.15934	0.02248	0.30010	0.11144	0.65982	0.00000	0.22092	0.03910	0.10557	0.01662						
D5S818	0.00000	0.41642	0.37243	0.00000	0.00000	0.19355	0.00000	0.00000	0.00293	0.00000	0.00000	0.00000	0.27761					
D7S820	0.01760	0.63832	0.21408	0.18573	0.06549	0.51222	0.82111	0.68426	0.28055	0.35973	0.48387	0.05181	0.14565	0.08407				
D9S938	0.00000	0.01173	0.05670	0.00000	0.00000	0.03715	0.00098	0.00000	0.11730	0.00000	0.00000	0.00000	0.03519	0.00000	0.43206			
SE33	0.33236	0.60117	0.14370	0.14370	0.07331	0.35875	0.56598	0.00098	0.29814	0.36559	0.07625	0.37634	0.18573	0.00000	0.36657	0.67058		
TH01	0.48680	0.27664	0.22776	0.54448	0.66080	0.05670	0.02444	0.15640	0.28739	0.12708	0.06843	0.03519	0.62072	0.14858	0.44770	0.04497	0.53568	

Supplementary Table 3

Supplemntary Table 3 - Allelic Frequencies part a.

Allele	D2S1338	D3S1358*	D3S2387	D3S2406	D5S818*	D5S2503	D7S820*	D9S938	D10S1237	D12S391	D13S317*	D16S539*	D16S753	D21S1437	D22S534	D22S689	SE33	TH01*
5					0.0002											0.0002		0.0022
6					0.0084											0.0116		0.1952
7					0.0058		0.0139				0.0003	0.0001				0.0055	0.0002	0.2717
7.1					0.0008													
8					0.0211		0.163		0.0016		0.082	0.0268			0.0006	0.0172	0.0001	0.1689
9					0.0306		0.1217		0.0006		0.0753	0.1811			0.0012	0.0825	0.0001	0.1623
9.3																		0.1958
10					0.0597		0.2909		0.0047		0.041	0.0868			0.0025	0.1028	0.001	0.0038
10.2																0.0005	0.0003	
11		0.0007			0.3159		0.2245		0.0056		0.2948	0.285			0.0015	0.2894	0.0024	0.0001
11.2																0.0067	0.0001	
12		0.0022			0.3567		0.1574		0.0016		0.3225	0.2469			0.0593	0.312	0.0055	
12.2																0.0046	0.0014	
13	0.0002	0.0049	0.0009		0.1869		0.0254		0.0001		0.133	0.1512			0.3023	0.1277	0.0127	
13.2																0.0009	0.0009	
14	0.0001	0.1015	0.0003		0.0128		0.0031		0.0003	0.0003	0.0503	0.0208			0.3886	0.0347	0.0323	
14.2																	0.0009	
15	0.0015	0.2914	0.0009		0.0011		0.0001		0.0023	0.0546	0.0008	0.001			0.1737	0.0037	0.055	
15.2			0.0267														0.002	
15.3										0.0001								
16	0.048	0.2805	0.0089		0.0001				0.0118	0.043					0.0648		0.0746	
16.1										0.0001								
16.2			0.0001														0.001	
16.3										0.0001								
17	0.1933	0.2162	0.0596						0.0922	0.1118					0.0052		0.083	
17.1										0.0003								
17.2			0.0023							0.0001							0.0004	
17.3										0.0065								
18	0.0738	0.094	0.0755						0.1082	0.2206					0.0003		0.1008	
18.1										0.0001								
18.2			0.0067														0.0003	
18.3										0.0078								
19	0.1271	0.0072	0.1031						0.1828	0.1553					0.0002		0.0973	
19.1										0.0009								
19.2			0.0039							0.0001							0.0001	
19.3										0.0033								
20	0.1297	0.0011	0.1099						0.244	0.1411							0.0648	
20.2			0.0244														0.0031	
21	0.0707	0.0002	0.134						0.1224	0.0897							0.0343	
21.2			0.0279														0.0164	
22	0.0809		0.1583						0.0874	0.0805							0.0121	
22.2			0.0161														0.0184	
23	0.1099		0.1283					0.001	0.0595	0.0529							0.002	
23.2			0.0126														0.0215	
24	0.082		0.0567					0.019	0.0417	0.0192							0.0005	
24.2			0.0041														0.0275	
25	0.0636		0.0273	0.0001				0.1557	0.0302	0.0101							0.0002	
25.2			0.0006														0.0336	
26	0.0167		0.0088	0.0002				0.2389	0.0029	0.0011							0.0001	
26.2			0.0008														0.0487	
27	0.002		0.0008	0.0001				0.2391	0.0003	0.0006							0.0004	
27.2																	0.0645	
28	0.0001		0.0002	0.003				0.1602	0.0002								0.0007	
28.2																	0.0563	
29				0.0135				0.1522									0.0005	
29.2																	0.0514	
30				0.0286				0.0318									0.0001	
30.2				0.001													0.0332	

Supplemntary Table 3 - Allelic Frequencies part b.

Allele	D2S1338	D3S1358*	D3S2387	D3S2406	D5S818*	D5S2503	D7S820*	D9S938	D10S1237	D12S391	D13S317*	D16S539*	D16S753	D21S1437	D22S534	D22S689	SE33	TH01*
31				0.0788				0.0019									0.0004	
31.2				0.0007													0.02	
32				0.1248				0.0005									0.0007	
32.2				0.0007													0.0099	
33				0.1286				0.0001									0.0003	
33.2																	0.003	
34				0.1303				0.0002										
34.2																	0.0011	
35				0.112													0.0002	
35.2																	0.0016	
36				0.102													0.0003	
36.2																	0.0003	
37				0.081														
38				0.0847														
39				0.054														
40				0.0267														
41				0.0154														
42				0.0096														
43				0.0032														
44				0.0011														
45				0.0003														
46				0.0001														
105														0.0009				
109														0.0197				
113														0.0438				
117														0.1716				
121														0.0899				
125														0.0929				
129														0.3283				
133														0.1287				
137														0.0967				
141														0.0251				
145														0.0023				
149														0.0001				
236													0.0001					
240													0.0016					
244													0.0152					
248													0.0173					
252													0.0514					
256													0.1819					
260													0.2442					
264													0.2077					
268													0.1803					
272													0.0678					
276													0.0224					
280													0.0084					
284													0.0018					
288													0.0002					
350						0.0032												
354						0.004												
358						0.0926												
362						0.1095												
366						0.3243												
370						0.3195												
374						0.1116												
378						0.0266												
382						0.0048												
386						0.0025												
390						0.0013												

*CODIS markers.

Supplementary Table 4

Supplentary Table 4 - Pairwise Fst [13] from 3 Brazilian populations: Bahia, Espírito Santo and Minas Gerais. ^{a}P=0.00000, ^{b}P=0.00000 and ^{c}P=0,32432.

	Bahia	Espírito Santo	Minas Gerais
Bahia	0.00000	+	+
Espírito Santo	0.00028^{a}	0.00000	-
Minas Gerais	0.00062^{b}	0.00002^{c}	0.00000

Supplementary Table 5

Suplentary Table 5 - Mutations observed at 18 STR loci in the populations from Bahia, Espírito Santo and Minas Gerais, Brazil.

Locus	Nº of meiosis	Nº of mutations	Mutation rate	95% confidence limits
D2S1338	31418	15	0.0005	0-0.0010
D3S1358	30944	11	0.0004	0-0.0007
D3S2387	12556	2	0.0002	0-0.0003
D3S2406	24641	24	0.0010	0-0.0019
D5S818	15941	10	0.0006	0-0.0013
D5S2503	27065	15	0.0006	0-0.0011
D7S820	34939	17	0.0005	0-0.0010
D9S938	30357	4	0.0001	0-0.0003
D10S1237	20082	11	0.0005	0-0.0011
D12S391	32769	50	0.0015	0-0.0031
D13S317	30974	18	0.0006	0-0.0012
D16S539	24195	11	0.0005	0-0.0009
D16S753	12438	3	0.0002	0-0.0005
D21S1437	23797	6	0.0003	0-0.0005
D22S534	23291	10	0.0004	0-0.0009
D22S689	18050	10	0.0006	0-0.0011
SE33	30821	68	0.0022	0-0.0044
TH01	27953	0	0	0-0

We estimated mutations rates for the 18 loci based on the analysis of the paternity cases (mother-son-father, son-father or son-mother). The rate was calculated using the number of mutations observed divided by the number of meiosis with the IC (95%).

3.2 - Capítulo II – *Genetic profile and admixture of the Brazilian population based in markers used for forensic applications*

Artigo submetido para publicação na *Forensic Science International – Genetics*.

A população brasileira é uma população miscigenada com contribuição de populações indígenas, europeias e africanas. Os objetivos deste trabalham eram: i) determinar o perfil genético das populações Brasileiras e ii) demonstrar a contribuição africana e europeia nas populações brasileiras. Neste estudo foram analisados 2.429 indivíduos não parentados extraídos do banco de dado 3 da Biocod Biotecnologia e 78 amostras do painel público disponibilizado pelo *Coriell Institute of Medical Research*, 24 indivíduos com ancestralidade africana, 31 europeus e 23 latino-americanos miscigenados (Hispânicos). Todos os indivíduos foram genotipados para os painéis de STR descritos por Pinto *et al* (2014). Os resultados demonstraram que os marcadores previamente caracterizados são informativos tanto para análises forenses quanto para estudos genético-populacionais. Todas as populações e marcadores estão em equilíbrio de Hardy-Weinberg após a correção de Bonferroni (*Supplementary Table 1-8*, *Table 1*) e são geneticamente diferentes (*Table 1*). As populações Brasileiras receberam uma maior contribuição europeia do que africana (*Table 2, Figure 2*). Nossos resultados mostram que a combinação estudada de 18 CODIS e não-CODIS loco é informativa para análise genética forense nas diversas regiões brasileiras, apesar de pequenas diferenças na estrutura da população, que são consistentes com a história demográfica brasileira dos últimos quinhentos anos.

GENETIC PROFILE AND ADMIXTURE OF THE BRAZILIAN POPULATION BASED ON MARKERS USED FOR FORENSIC APPLICATIONS

Laélia Maria Pinto[1,2], Fernanda SG Kehdy[2] , Camila Coutinho Bernardes[1], Cristiane Lommez de Oliveira[3], Luciana Lara dos Santos[4], Eduardo Tarazona-Santos[2]

[1]Hermes Pardini. Avenida das Nações, 2448, Vespasiano, MG, zip code 33200-000, Brazil.

[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Av. Antonio Carlos 6627, Pampulha. Post office box 486, Belo Horizonte, MG, zip code 31270-910, Brazil.

[3]Biocod Biotecnologia. Rua Aimorés, 66, 6th floor, Funcionários, Belo Horizonte, MG, zip code 30140-070, Brazil.

[4]Universidade Federal de São João Del Rei – Campus Centro-oeste Dona Lindu. Rua Sebastião Gonçalves Coelho, 400, Divinópolis, MG, zip code 35501-296, Brazil.

CORRESPONDING AUTHOR:

Laélia Maria Pinto

Avenida das Nações, 2448

Vespasiano, MG, zip code 33200-000, Brazil

Phone: 55 31 36726993, Fax: 55 31 36294873

Email: laeliap@hotmail.com

Abstract

Brazilians trace their origins to the original Amerindians and two main sources of immigration: Africans and Europeans. Based on 18 STR used in forensic applications, we study the admixture and genetic structure of the Brazilian population. We analyze 2,429 unrelated individuals extracted to the Biocod's database classified in in five Brazilian geographic regions, and 78 ethnically diverse individuals with European and African ancestry. The distribution of allelic frequencies across the five Brazilian populations shows significant differences in 13 loci. All markers are highly informative for forensic purposes. The five analyzed Brazilian regional populations (North, Northeast, Midwest, Southeast and South) fit the Hardy -Weinberg model for all loci, with low level of genetic structure between them, mainly determined by differences in the levels of African vs. European continental ancestry

Keywords: Brazilian population, STR, Forensic, ancestry, AMOVA, Structure.

1. Introduction

Historically, the Brazilian population always experienced high levels of intermarriage between ethnic groups, and Brazilians are known to be intensively admixed with Amerindian, European and African ancestries and followed variable patterns of multidirectional introgression according to the social and historical conditions in each geopolitical region along the last five centuries up to the present day [1-3]. In this study we assess the genetic structure among the five Brazilian macro-region using data from eighteen Short Tandem Repeat (STR) loci used for forensic genetics purposes [4], genotyped in a large number of individuals.

2. Materials and Methods

2.1. Population samples

A STR database was built based on individuals of all Brazilian states who participated in paternity tests conducted in the Biocod Biotechnology laboratory. In this study we analyze 2,429 unrelated individuals extracted from the Biocod's database. These individuals were divided according to the geographic region of birth-place (Figure 1): North (N; 230 individuals), Northeast (NE; 989 individuals); Midwest (MD; 36 individuals); South (S; 175 individuals) and Southeast (SE; 999 individuals). Brazilian populations were compared with a publicly available panel that includes 24 individuals of African ancestry, 31 Europeans and 23 admixed Latin Americans (i.e., Hispanics) from the Coriell Cell Repositories (Coriell Institute of Medical Research, Camden, NJ, USA).

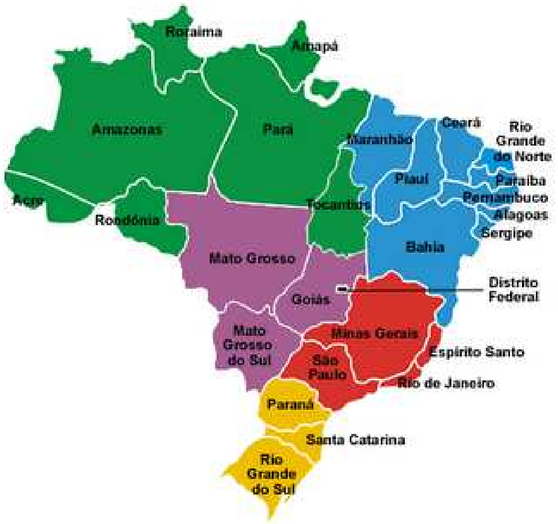


Figure 1 – Map of Brazil divided accord to the geographic regions. The individuals were divided according to the geographic region of birth-place: North (Green; N; 230 individuals), Northeast (Blue;

NE; 989 individuals); Midwest (Violet; MD; 36 individuals); South (Yellow; S; 175 individuals) and Southeast (Red; SE; 999 individuals).

2.2. STR amplification

All the 2,429 Brazilian samples and the 78 reference samples were genotyped for two STR-multiplex (Panel 1 and Panel 2), for a total of 18 loco used in the Biocod lab routine. These markers are divided in two groups: CODIS markers - D3S1358 (Panel 1), D5S818 (Panel 2), D7S820 (Panel 1), D13S317 (Panel 1), D16S539 (Panel 1) and TH01 (Panel 1); and non-CODIS markers - D2S1338 (Panel 2), D3S2387 (Panel 2), D3S2406 (Panel 2), D5S2503 (Panel 2), D9S938 (Panel 1), D10S1237 (Panel 1), D12S391 (Panel 2), D16S753 (Panel 2), D21S1437 (Panel 1), D22S534 (Panel 2), D22S689 (Panels 1 and 2) and SE33 (Panel 1). Data for both Panels 1 and 2 were genotyped as detailed by Pinto et al. [4].

2.3. Allele frequencies and population genetics statistics

Allele frequencies were calculated using GENEPOP [5] for each population. The statistics (MP, matching probability; PIC, polymorphic information content; PD, power of discrimination; PE, power of exclusion; TPI, typical paternity index), that estimate the informativeness of the markers, were calculated using the PowerStats program v1.2.xls (Promega Corporation®). Observed and expected heterozygosity [6] as well as deviation from Hardy–Weinberg equilibrium (HWE; heterozygote deficiency) were estimated using the Arlequin software [7,8,9].

The Analyses of Molecular Variance (AMOVA) was carried out on the dataset by using the Arlequin 3.5 software [9]. The analysis included data for all eight populations (North, Northeast, Midwest, South and Southeast – Brazilian populations; African ancestry; Europeans and Hispanic).

The pairwise population genetic distance, F_{ST}, was estimated according to Slatkins [10] by using the program Arlequin 3.5 [9]. The significance of F_{ST} was determined using permutation tests (1000 permutations) and 0.05 significance level.

Population structure was also analyzed using the Bayesian model-based analysis implemented in the software STRUCTURE 2.3.4 [11]. Because we did not have data for Native American samples, and several studies suggest that at least in the Northeast, Southeast and South of Brazil, the Native American contribution is low [12], we assumed two parental populations (K=2). We performed five independent runs of STRUCTURE, with 100,000 repetitions and a burn-in period of 20,000 following the admixture model and correlated allele frequencies, and considering the African ancestry and Europeans individuals as belonging to parental populations and the Hispanics and Brazilian individuals as admixed.

3. Results and discussion

The Supplementary Tables 1-8 show the allele frequencies, forensic genetics statistics and exact tests of Hardy-Weinberg equilibrium for each locus and population. All populations and loco are in Hardy-Weinberg equilibrium after Bonferroni correction (P = 0.00034). All markers showed a high degree of genetic polymorphism, PIC values were higher than 0.6 (60%). Also, the values of PIC were higher than 0.5 what indicates this STR system to be informative and useful for identification purposes [13,14].

Table 1- Locus by locus F-statistics: F_{IS}, F_{ST} and F_{IT}

Locus	F_{IS} (P value)	F_{ST} (P value)	F_{IT} (P value)
D10S1237	0.01841 (0.01822)	0.00232 (0.02475)	0.02068 (0.00980)
D12S391	0.00329 (0.33782)	0.00112 (0.39139)	0.00440 (0.30000)
D13S317	0.02427 (0.00891)	0.00013 (0.89792)	0.02440 (0.00941)
D16S539	-0.00253 (0.59634)	0.00065 (0.78426)	-0.00188 (0.58733)
D16S753	0.02557 (0.01881)	0.00067 (0.90921)	0.02622 (0.01733)
D21S1437	0.00149 (0.44644)	0.00189 (0.05436)	0.00338 (0.38594)
D22S534	0.01045 (0.19950)	0.00316 (0.00931)	0.01357 (0.16099)
D22S689	-0.01695 (0.92733)	0.00312 (0.01921)	-0.01377 (0.89604)
D2S1338	0.00696 (0.16861)	0.00121 (0.30653)	0.00816 (0.14624)
D3S1358	0.00946 (0.20089)	0.00112 (0.45089)	0.01056 (0.17505)
D3S2387	0.01840 (0.02119)	0.00213 (0.21564)	0.02049 (0.01238)
D3S2406	0.02062 (0.00208)	0.00066 (0.92733)	0.02127 (0.00139)
D5S2503	0.01233 (0.13089)	0.00240 (0.04069)	0.01470 (0.10218)
D5S818	0.01354 (0.20347)	0.00351 (0.11436)	0.01700 (0.17119)
D7S820	0.01464 (0.07396)	0.00021 (0.95832)	0.01485 (0.06584)
D9S938	0.02473 (0.00673)	**0.00422 (0.00000)**	0.02885 (0.00228)
SE33	0.00454 (0.18376)	0.00107 (0.43970)	0.00561 (0.15554)
THO1	0.01672 (0.05356)	**0.00393 (0.00010)**	0.02059 (0.02683)

P = 0.00034, after Bonferroni correction. Significant P values are highlighted.

The AMOVA results showed low variation among populations for the studied loci (Table 1). Fst values range from 0.00013 to 0.00422 for D13S317 and D9S938 respectively. Normally, it is expected that F_{ST} values between populations are around 0.05 [15]. The F_{IT} and F_{IS} did not show significant values considering the Bonferroni correction, consistently with the Hardy-Weinberg equilibrium test. In general, Fst analysis (Table 2) shows that the studied Brazilian populations are closer to the Europeans and Latin American/Hispanic populations than to the African ancestry Coriell individuals.

Table 2 - Pairwise FST Genetic Distance between Populations

	AFR	EUR	HIS	N	NE	MD	S	SE
AFR		+	+	+	+	+	+	+
EUR	0.02454		+	+	+	-	-	+
HIS	0.02069	0.01218		-	-	-	-	-
N	0.01368	0.00768	0.00164		+	-	+	-
NE	0.01534	0.00401	0.00376	0.00145		-	+	-
MD	0.01220	0.00462	0.00458	0.00000	0.00000		-	-
S	0.02705	0.00204	0.00347	0.00327	0.00243	0.00121		+
SE	0.01361	0.00392	0.00409	0.00157	0.00011	0.00073	0.00307	

The significant values were represented by "+" signal. Abbreviations: AFR - African ancestry; EUR - Europeans; HIS - Latin Americans/Hispanicos; N - North; NE - Northeast; MD - Midwest; S - South; and SE - Southeast.

These results are confirmed by the STRUCTURE analysis (Figure 2), that suggests that African ancestry contributes between 17-23% to the studied Brazilian populations. Among Brazilian populations; Midwest, Southeast, North and Northeast populations are nearest to and received more admixture from the African population than the Southern population. The observed results are consistent with the demographic history of the Brazilian population [1,2]. This result has the limitation of not being including a Native American ancestry population in the analyses (i.e. the third continental ancestral component of Brazilians) due to the lack of this data. However, because Native American ancestry tends to be low in urban Brazilian populations [16], the absence of this data should not critically affect the observed trend in admixture.

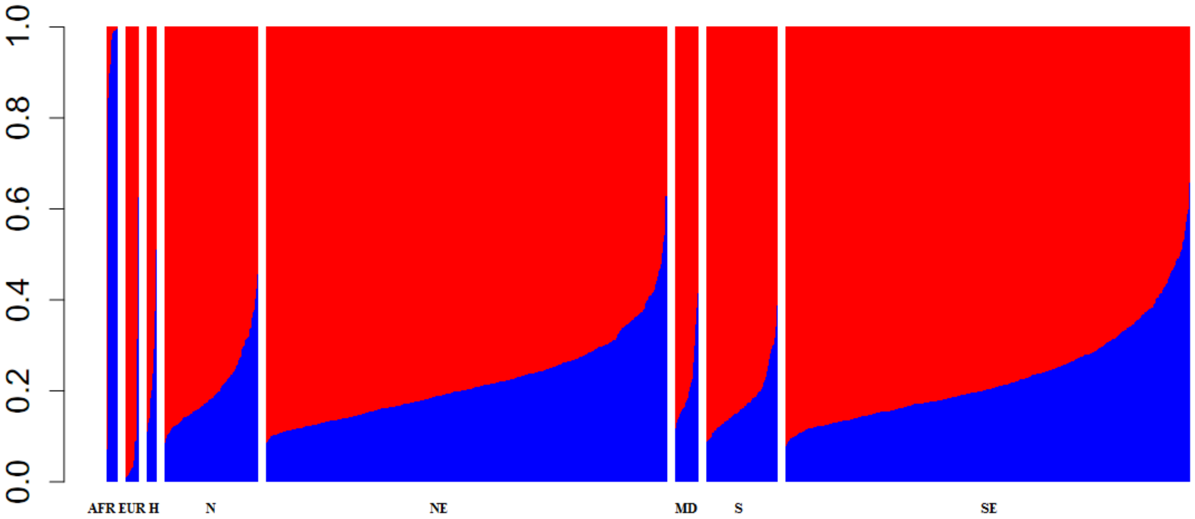


Figure 2 – Barplot of European (red) and African (blue) individual admixture inferred by Structure assuming two parental populations: AFR (African ancestry) and EUR (Europeans) from the Coriell repository. Admixed population are Latin American/Hispanics (H) from the Coriell repository and Brazilians (N: North, NE: Northeast, MD: Midwest, S: South and SE: Southeast). Estimated mean

proportions across individuals of European and African admixture are shown for Brazilian populations (N: 20.4% AFR and 79.6% EUR; NE: 21.8% AFR and 78.2% EUR; MD: 20.1% AFR and 79.9% EUR; S: 17.4% AFR and 82.6% EUR; SE: 23.2% AFR and 76.8% EUR) and Latin American/Hispanic (H: 21.9% AFR and 78.1% EUR)

In conclusion, our results show that the studied combination of 18 CODIS and non-CODIS loco are informative for forensic genetic analysis across the different Brazilian regions, despite small differences in population structure, which are consistent with the Brazilian demographic history of the last five-hundred years.

Acknowledgments

This publication was supported by the Institute Hermes Pardini. We would like to thank the Department Genetics Coordinator of Institute Hermes Pardini Amanda Mafia de Castro and Core Lab Manager of Institute Hermes Pardini Vanessa Oliveira.

References

[1] Giolo S. R., Soler J. M. P., Greenway S. C., Almeida M.A.A., Andrade M., Seidman J. G., Seidman C. E., Krieger J. E., Pereira A. C., Brazilian urban population genetic structure reveals a high degree of admixture. **European Journal of Human Genetics**, 20, 2012,111-116.

[2] Repnikova E. A., Rosenfeld J. A., Bailes A., Weber C., Erdman L., McKinney A., Ramsey S., Hashimoto S., Thrush D. L., Astbury C., Reshmi S. C., Shaffer L. G., Gastier-Foster J. M., Pyatt R. E. Characterization of copy number variation in genomic regions containing STR loci using array comparative genomic hybridization. **Forensic Science International: Genetics**, 7, 2013, 475-481.

[3] Lins T. C., Vieira R. G., Abreu B. S., Grattapaglia D., Pereira R. W. Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. **American Journal of Human Biology**, 22, 2010,187-192.

[4] Pinto L. M.., Oliveira C. L., Santos L. L., Tarazona-Santos E. Molecular characterization and population genetics of non-CODIS microsatellites used for forensic applications in Brazilian populations. **Forensic Science International: Genetics**, 9, 2014, e16-e17.

[5] Raymond M., Rousset F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. **Journal of Heredity**, 86, 1995, 248-249.

[6] Nei, M. **Molecular Evolutionary Genetics**. New York: Columbia University Press, 1987.

[7] Excoffier L., Estoup A., Cornuet J. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. **Genetics,** 169, 2005, 1727-1738.

[8] Excoffier L., Hofer T., Foll M. Detecting loci under selection in a hierarchically structured population. **Heredity**, 103, 2009, 285-298.

[9] 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,** 10, 2010, 564–567.

[10] Slatkin M., Voelm L. F_{ST} in a hierarchical island model. **Genetics**, 127, 1991, 627-629.

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

[12] Pena S.D.J., Di Pietro G., Fuchshuber-Moraes M., Genro J. P., Hutz M. H., Kehdy F. S. G., Kohlrausch F., Magno L. A. V., Montenegro R. C., Moraes M. O., Moraes M. E. A., Moraes M. R., Ojopi E. B., Perini J. A., Racciopi C., Ribeiro-dos-Santos A. K. C., Rios-Santos F., Romano-Silva M. A., Sortica V. A., Suarez-Kurtz G. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. **PlosOne**, 6, 2, 2011, 1-9.

[13] Aguiar V. R. C., Wolfgramm E. V., Malta F. S. V., Bosque A. G., Mafia A. C., Almeida V. C. O., Caxito F. A., Pardini V. C., Ferreira A. C. S., Louro I. D. Updated Brazilian STR allele frequency data using over 100,000 individuals: an analysis of CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA loci, **Forensic Science International: Genetics**, 6, 2012, 504-509.

[14] Manamperi A., Hapuarachchi C., Gunawardene N. S., Bandara A., Dayanath D., Abeyewickreme W. STR polymorphisms in Sri Lanka: evaluation of forensic utility in identification of individuals and parentage testing. **Ceylon Medical Journal**. 54, 3, 2009, 85-89.

[15] Holsinger K. E., Weir B. S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. **Nature Reviews Genetics,** 10, 2009, 639-650.

[16] Manta F. S. N., Pereira R. Vianna R., Araújo A. R. B., Gitaí D. L. G., Silva D. A., Wolfgramm E. V., Pontes I. M., Aguiar J. I., Moraes M. O., Carvalho E. F. C., Gusmão L. Revisiting the genetic ancestry of brazilians using autosomal AIM-Indels. **PlosOne**, 8, 9, 2013, e75145.

Supplementary Table 1

Supplementary Table 1 - Allele frequencies of seventeen autosomal STR loci in African ancestry part a

Alelo	D10S1237	D12S391	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																	
5			0.0652														0.0208
6							0.0417										0.1458
7																0.1042	0.4583
7.1																	
8							0.0208						0.1042	0.2083			0.2292
9			0.2391			0.0208	0.1042							0.1042			0.0417
9.3																	0.0417
10			0.0870			0.0208	0.1250						0.0417	0.2708			0.0208
10.2																	
11			0.3261				0.0625						0.1667	0.3542			0.0417
11.2	0.0435						0.0208										
12			0.1304			0.0833	0.5000						0.2292	0.0625			
12.2																	
13			0.1522			0.1458	0.1042						0.4375				
13.2																	
14						0.3333	0.0208		0.0435				0.0208			0.0208	
14.2																	
15		0.0833				0.3542			0.3043							0.0417	
15.2																	
16		0.0625				0.0417			0.3043	0.0417						0.0625	
16.1																	
16.2										0.0208							
17	0.2609	0.1458						0.1250	0.3261	0.0208						0.1042	
17.2																0.0208	
17.3																	
18	0.1087	0.2500						0.0208	0.0217	0.1250						0.1250	
18.2																	
18.3																	
19	0.1087	0.0625						0.2708		0.1458						0.1458	
19.2										0.0208						0.0208	
20	0.2174	0.0833						0.1667		0.1458						0.0417	
20.2																	
21	0.0870	0.0833						0.0833		0.0625						0.0417	
21.2										0.1250						0.0208	
22	0.1087	0.0833						0.1042		0.0417							
22.2										0.0208						0.0417	
23	0.0435	0.0625						0.1250		0.0625							
23.2										0.0208							
24	0.0217	0.0208						0.0625		0.0625					0.0227		
24.2										0.0208							
25		0.0208						0.0208							0.1818		
25.2		0.0208														0.0208	
26								0.0208							0.2045		
26.2		0.0208														0.0417	
27															0.2500		
27.2											0.0417					0.0417	
28										0.0208					0.1818		
28.2											0.0208					0.0833	
29										0.0417					0.1591		
29.2											0.0625						
30																	
30.2																0.0208	
30.3																	
31											0.0625						
31.2											0.0208						
32											0.0417						
32.2																	
33											0.1667						
33.2																	
34											0.1250						
34.2																	
35											0.0417						
35.2																	
36											0.0833						
36.2																	
37											0.0625						
37.2											0.0208						
38											0.0833						
39											0.1250						
40																	
41											0.0417						
42																	
43																	
44																	
45																	

Supplementary Table 1 - Allele frequencies of seventeen autosomal STR loci in African ancestry part b

Alelo	D10S1237	D12S391	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105					0.0208												
109					0.0625												
113					0.1042												
117					0.1042												
121					0.1042												
125					0.2292												
129					0.2917												
133					0.0833												
137																	
141																	
145																	
244				0.0208													
248				0.0833													
252				0.1458													
256				0.2083													
260				0.1667													
264				0.1250													
268				0.1042													
272				0.0625													
276				0.0833													
280																	
284																	
288																	
350												0.0208					
354																	
358												0.0417					
362												0.0833					
366												0.3333					
370												0.4167					
374												0.0625					
378												0.0417					
382																	
386																	
390																	
N	23	24	23	24	24	24	24	24	23	24	24	24	24	24	22	24	24
OH (%)	0.782	0.917	0.826	0.792	0.708	0.792	0.792	0.792	0.609	0.833	0.958	0.750	0.625	0.708	0.682	0.792	0.792
EH (%)	0.856	0.894	0.802	0.881	0.836	0.748	0.721	0.863	0.722	0.926	0.926	0.715	0.731	0.759	0.822	0.937	0.725
P	0.662	0.316	0.902	0.225	0.071	0.738	0.991	0.074	0.786	0.067	0.675	0.565	0.043	0.396	0.226	0.086	0.529
MP	0.0662	0.069	0.096	0.066	0.087	0.142	0.108	0.080	0.134	0.056	0.052	0.167	0.160	0.132	0.095	0.047	0.153
Exp. as 1 in	15.114	14.400	10.373	15.158	11.520	7.024	9.290	12.522	7.451	18.000	19.200	6.000	6.261	7.579	10.522	21.160	6.545
PIC	0.818	0.864	0.754	0.848	0.796	0.690	0.682	0.827	0.648	0.899	0.900	0.653	0.674	0.700	0.774	0.909	0.673
PD	0.934	0.931	0.904	0.934	0.913	0.858	0.892	0.920	0.866	0.944	0.948	0.833	0.840	0.868	0.905	0.953	0.847
PE	0.567	0.830	0.648	0.584	0.441	0.584	0.584	0.584	0.301	0.662	0.915	0.510	0.322	0.441	0.401	0.567	0.584
TPI	2.300	6.000	2.875	2.400	1.714	2.400	2.400	2.400	1.278	3.000	12.000	2.000	1.330	1.714	1.571	2.300	2.400

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 2

Supplementary Table 2 - Allele frequencies of seventeen autosomal STR loci in European population part a

Alelo	D10S1237	D12S391	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																	
5																	0.0161
6																	0.1774
7							0.0156										0.2097
7.1																	
8							0.0156						0.0156	0.1875			0.0968
9			0.1667				0.0469						0.0781	0.1406			0.2258
9.3																	0.2581
10			0.0500				0.0938						0.0938	0.2656			0.0161
10.2																	
11			0.3167				0.3594		0.0156				0.3281	0.1719			
11.2																	
12			0.2333			0.0312	0.3125						0.2344	0.1719			
12.2																	
13			0.1667			0.1875	0.1094		0.0156				0.2500	0.0625			
13.2																	
14			0.0667			0.5312	0.0312		0.0312							0.0312	
14.2																	
15		0.0469				0.1094	0.0156		0.2188							0.0156	
15.2																	
16						0.1406		0.0312	0.3594							0.0469	
16.1																	
16.2																	
17	0.0469	0.1094						0.2344	0.2500	0.0312						0.0781	
17.1																0.0156	
17.2																	
17.3		0.0156															
18	0.0312	0.1562						0.1250	0.1094	0.0469						0.0625	
18.2																0.0312	
18.3																	
19	0.2969	0.1406						0.2031		0.0312						0.1094	
19.2		0.0156															
20	0.2500	0.1406						0.1250		0.0938						0.0312	
20.2																0.0156	
21	0.0938	0.1094						0.0469		0.1875						0.0312	
21.2																0.0156	
22	0.1094	0.0469						0.0156		0.2969						0.0156	
22.2																0.0625	
23	0.0625	0.1406						0.0781		0.2188							
23.2																0.0781	
24	0.0781	0.0312						0.0625		0.0781					0.0294		
24.2																0.0469	
25	0.0312	0.0469						0.0625							0.0882		
25.2																0.0312	
26								0.0156							0.1176		
26.2																0.0312	
27															0.3235		
27.2																0.0312	
28										0.0156					0.2353		
28.2																0.0469	
29															0.2059		
29.2											0.0750					0.0781	
30																	
30.2																0.0156	
30.3																	
31											0.0750						
31.2																0.0156	
32											0.0500						
32.2																0.0156	
33											0.1750					0.0156	
33.2																	
34											0.0750						
34.2																	
35											0.1250						
35.2																	
36											0.0750						
36.2																	
37											0.1250						
38											0.1500						
39																	
40											0.0500						
41											0.0250						
42																	
43																	
44																	
45																	

Supplementary Table 2 - Allele frequencies of seventeen autosomal STR loci in European population part b

Alelo	D10S1237	D12S391	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105																	
109																	
113																	
117					0.0781												
121					0.0312												
125					0.1250												
129					0.4844												
133					0.0469												
137					0.1562												
141					0.0781												
145																	
248				0.0312													
252				0.1250													
256				0.0781													
260				0.2344													
264				0.2031													
268				0.2656													
272				0.0625													
276																	
280																	
284																	
288																	
350												0.0156					
354																	
358												0.0625					
362												0.0781					
366												0.3750					
370												0.2031					
374												0.2031					
378												0.0625					
382																	
386																	
390																	
N	32	32	30	32	32	32	32	32	32	32	20	32	32	32	17	32	31
OH (%)	0.969	0.906	0.733	0.813	0.719	0.500	0.875	0.844	0.750	0.813	1.000	0.750	0.781	0.688	0.824	0,938	0.806
EH (%)	0.827	0.898	0.796	0.819	0.721	0.660	0.760	0.869	0.759	0.822	0.910	0.775	0.772	0.824	0.799	0,959	0.810
P	0.037	0.334	0.298	0.262	0.585	0.090	0.641	0.136	0.692	0.872	0.982	0.174	0.456	0.324	0.801	0.449	0.857
MP	0.125	0.053	0.102	0.094	0.127	0.170	0.131	0.068	0.117	0.072	0.060	0.113	0.115	0.072	0.114	0.033	0.084
Exp. as 1 in	8.000	18.963	9.783	10.667	7.877	5.885	7.642	14.629	8.533	13.838	16.667	8.828	8.678	13.838	8.758	30.118	11.864
PIC	0.792	0.873	0.751	0.779	0.682	0.609	0.712	0.839	0.706	0.785	0.877	0.730	0.721	0.784	0.741	0.942	0.766
PD	0.875	0.947	0.898	0.906	0.873	0.830	0.869	0.932	0.883	0.928	0.940	0.887	0.885	0.928	0.886	0.967	0.916
PE	0.937	0.808	0.482	0.622	0.458	0.188	0.745	0.683	0.510	0.622	0.898	0.510	0.565	0.409	0.643	0.872	0.611
TPI	16.000	5.333	1.875	2.667	1.778	1.000	4.000	3.200	2.000	2.667	10.000	2.000	2.286	1.600	2.833	8.000	2.583

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 3

Supplementary Table 3 - Allele frequencies of seventeen autosomal STR loci in Latin American/Hispanic population part a

Alelo	D10S1237	D12S391	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																	
5																	
6																	0.2273
7													0.0217				0.2273
7.1																	
8			0.0217				0.0217							0.1364			0.0909
9			0.0870				0.0435						0.0652	0.1136			0.1591
9.3																	0.2500
10			0.1739				0.1087						0.0870	0.2727			0.0227
10.2																	
11			0.2826				0.3261		0.0217				0.3913	0.3409			0.0227
11.2	0.0227																
12			0.1957			0.0435	0.3913						0.2391	0.1136			
12.2																	
13			0.2174			0.4130	0.0652						0.1957	0.0227		0.0227	
13.2																	
14			0.0217			0.3696	0.0435		0.0435								
14.2																	
15	0.0227	0.1087				0.0652			0.3043							0.0227	
15.2																0.0227	
16		0.0435				0.1087		0.0909	0.2391	0.0435						0.0227	
16.1																	
16.2										0.0217							
16.3										0.0217							
17	0.1364	0.0870						0.1818	0.2174	0.0870						0.0909	
17.2																	
17.3		0.0435															
18	0.1364	0.1957						0.0227	0.1522	0.0652						0.1136	
18.2																0.0227	
18.3		0.0435															
19	0.2273	0.2609						0.0909	0.0217	0.1957						0.1136	
19.2																	
20	0.2273	0.0217						0.1364		0.0652						0.0682	
20.2										0.0217							
21	0.0455	0.0870						0.0455		0.1522						0.0227	
21.2										0.0217							
22	0.0455	0.0652						0.0227		0.1087							
22.2																	
23	0.0682	0.0217						0.2727		0.1087							
23.2																0.0227	
24	0.0227	0.0217						0.0227		0.0652							
24.2																	
25	0.0227							0.0682							0.1842		
25.2																	
26	0.0227							0.0455							0.2895		
26.2																0.1591	
27															0.1579		
27.2																0.0909	
28															0.2368		
28.2																0.0455	
29										0.0217					0.1316		
29.2																0.0227	
30																	
30.2																0.0909	
30.3																	
31											0.1087						
31.2																	
32											0.0217						
32.2																0.0455	
33											0.1739						
33.2																	
34											0.0652						
34.2																	
35											0.0870						
35.2																	
36											0.0652						
36.2																	
37											0.1522						
38											0.1304						
39											0.0435						
40											0.0870						
41											0.0435						
42																	
43																	
44																	
45											0.0217						

Supplementary Table 3 - Allele frequencies of seventeen autosomal STR loci in Latin American/Hispanic population part b

Alelo	D10S1237	D12S391	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105																	
109																	
113																	
117					0.2273												
121					0.0455												
125					0.0909												
129					0.2273												
133					0.1591												
137					0.1818												
141					0.0682												
145																	
244				0.0435													
248																	
252				0.0217													
256				0.0870													
260				0.3043													
264				0.3043													
268				0.1739													
272				0.0435													
276				0.0217													
280																	
284																	
288																	
350																	
354																	
358												0.1087					
362												0.0435					
366												0.1957					
370												0.4348					
374												0.1087					
378												0.1087					
382																	
386																	
390																	
N	22	23	23	23	22	23	23	22	23	23	23	23	23	22	19	22	22
OH (%)	0.909	0.913	0.783	0.870	0.864	0.652	0.870	0.818	0.652	0.870	0.826	0.783	0.739	0.591	0.684	0.818	0.955
EH (%)	0.868	0.874	0.814	0.789	0.842	0.690	0.736	0.867	0.794	0.910	0.909	0.752	0.756	0.782	0.805	0.932	0.818
P	0.629	0.404	0.248	0.590	0.703	0.455	0.571	0.629	0.155	0.107	0.189	0.311	0.106	0.143	0.528	0.425	0.466
MP	0.079	0.074	0.108	0.127	0.087	0.161	0.191	0.066	0.108	0.062	0.059	0.142	0.149	0.112	0.097	0.045	0.120
Exp. as 1 in	12.737	13.564	9.281	7.896	11.524	6.224	5.238	15.125	9.281	16.030	17.065	7.053	6.696	8.963	10.314	22.000	8.345
PIC	0.831	0.841	0.765	0.738	0.799	0.617	0.677	0.831	0.742	0.881	0.879	0.703	0.699	0.729	0.749	0.904	0.769
PD	0.921	0.926	0.892	0.873	0.913	0.839	0.809	0.934	0.892	0.938	0.941	0.858	0.851	0.888	0.903	0.955	0.880
PE	0.814	0.822	0.567	0.734	0.722	0.358	0.734	0.633	0.358	0.734	0.648	0.567	0.491	0.280	0.404	0.633	0.908
TPI	5.500	5.750	2.300	3.833	3.667	1.438	3.833	2.750	1.438	3.833	2.875	2.300	1.917	1.222	1.583	2.750	11.000

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 4

Supplementary Table - 4 Allele frequencie of eighteen autosomal STR loci in North population of Brazil part a

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																		
5														0.0036				
6								0.0119						0.0144				0.2303
7			0.0023					0.0030						0.0216	0.0133			0.3048
7.1																		
8			0.0787	0.0232				0.0208						0.0072	0.1327			0.1316
9			0.1065	0.1598			0.0025	0.0685						0.0360	0.1040			0.1009
9.3																		0.2237
10			0.0440	0.0902				0.0893						0.0647	0.2832			0.0088
10.2																		
11			0.2569	0.3067			0.0025	0.3274						0.3669	0.2832			
11.2								0.0030									0.0022	
12			0.3009	0.2448			0.0225	0.2917						0.3237	0.1416		0.0044	
12.2																		
13			0.1435	0.1495			0.3775	0.1548		0.0049				0.1547	0.0376		0.0133	
13.2																	0.0022	
14			0.0671	0.0258			0.3975	0.0238		0.0711				0.0072	0.0044		0.0156	
14.2																		
15	0.0055	0.0239					0.1300	0.0060		0.3505							0.0489	
15.2											0.0191							
16	0.0083	0.0261					0.0650		0.0459	0.3186	0.0127						0.1044	
16.1																		
16.2																	0.0022	
17	0.0442	0.0935					0.0025		0.2271	0.1691	0.1274						0.0800	
17.2																		
17.3		0.0087																
18	0.1077	0.2326							0.0677	0.0760	0.0764						0.1133	
18.2																	0.0022	
18.3		0.0043																
19	0.1823	0.2022							0.1463	0.0098	0.0860						0.1044	
19.2											0.0032							
20	0.2182	0.1783							0.1179		0.0796						0.0644	
20.2											0.0159							
21	0.1188	0.0804							0.0764		0.1210						0.0200	
21.2											0.0064						0.0133	
22	0.1077	0.0804							0.0677		0.1624						0.0111	
22.2											0.0096						0.0156	
23	0.0801	0.0435							0.1048		0.1497							
23.2											0.0096						0.0133	
24	0.0663	0.0152							0.0568		0.1497					0.0154		
24.2											0.0096						0.0356	
25	0.0608	0.0065							0.0677							0.2829		
25.2																	0.0356	
26		0.0065							0.0218							0.1908	0.0022	
26.2																	0.0556	
27												0.0028				0.2083		
27.2																	0.0600	
28											0.0350					0.1689		
28.2																	0.0600	
29											0.0096	0.0028				0.1162		
29.2																	0.0422	
30												0.0284				0.0175		
30.2																	0.0444	
30.3																		
31												0.0852						
31.2																	0.0133	
32												0.0909						
32.2																	0.0133	
33												0.1222						
33.2																	0.0044	
34												0.1222						
34.2																		
35												0.0938						
35.2																	0.0022	
36												0.1136						
36.2																		
37												0.0994						
38												0.1108						
39												0.0739						
40												0.0341						
41												0.0057						
42												0.0142						
43																		
44																		
45																		

Supplementary Table - 4 Allele frequencie of eighteen autosomal STR loci in North population of Brazil part b

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105																		
109						0.0089												
113						0.0356												
117						0.1467												
121						0.0978												
125						0.0956												
129						0.2911												
133						0.1356												
137						0.1533												
141						0.0333												
145						0.0022												
248					0.0160													
252					0.0256													
256					0.1699													
260					0.3045													
264					0.2083													
268					0.1891													
272					0.0545													
276					0.0256													
280					0.0064													
284																		
288																		
350													0.0025					
354																		
358													0.1465					
362													0.1692					
366													0.2904					
370													0.2601					
374													0.1010					
378													0.0202					
382													0.0051					
386													0.0051					
390																		
N	181	230	216	194	156	225	200	168	229	204	157	176	198	139	126	228	225	228
OH (%)	0.895	0.861	0.792	0.778	0.776	0.876	0.585	0.786	0.904	0.765	0.873	0.892	0.753	0.712	0.80531	0.781	0.933	0.754
EH (%)	0.868	0.850	0.801	0.7910	0.797	0.833	0.680	0.772	0.879	0.738	0.897	0.906	0.789	0.733	0.79127	0.799	0.936	0.778
P	0.494	0.772	0.566	0.232	0.457	0.711	0.010	0.762	0.672	0.051	0.435	0.111	0.182	0.900	0.88750	0.604	0.096	0.703
MP	0.037	0.039	0.068	0.069	0.066	0.055	0.275	0.095	0.025	0.106	0.026	0.023	0.076	0.116	0,0726084	0.069	0.013	0.082
Exp. as 1 in	27.369	25.862	14.809	14.448	15.161	18.289	3.640	10.561	40.762	9.434	38.940	44.001	13.237	8.609	13,772502	14.447	78.242	12.191
PIC	0.851	0.820	0.765	0.749	0.762	0.809	0.612	0.730	0.865	0.680	0.884	0.895	0.751	0.671	0,7505398	0.768	0.932	0.746
PD	0.963	0.961	0.932	0.931	0.934	0.945	0.725	0.905	0.975	0.894	0.974	0.977	0.924	0.884	0,9273916	0.931	0.987	0.918
PE	0.785	0.734	0.594	0.586	0.574	0.744	0.146	0.566	0.826	0.564	0.740	0.780	0.527	0.434	0,6244437	0.573	0.867	0.526
TPI	4.763	3.844	2.467	2.413	2.343	3.982	0.907	2.292	5.864	2.281	3.925	4.658	2.082	1.688	2,6818182	2.335	7.700	2.076

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 5

Supplementary Table 5 - Allele frequencies of eighteen autosomal STR in Northeast population of Brazil part a

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																		0.0005
5																		0.0005
6								0.0113						0.0056				0.2158
7								0.0077						0.0182	0.0175			0.2503
7.1														0.0014				
8			0.0888	0.0228				0.0246						0.0084	0.1638		0.0005	0.1591
9	0.0006		0.0836	0.1808			0.0038	0.0584						0.0238	0.1155			0.1648
9.3																		0.1998
10	0.0012		0.0491	0.1023			0.0016	0.0928						0.0532	0.2700			0.0093
10.2								0.0007										
11	0.0075		0.2847	0.2870				0.3165		0.0005				0.3235	0.2428		0.0015	
11.2								0.0021									0.0010	
12	0.0052		0.3161	0.2492			0.0638	0.3418		0.0042				0.3683	0.1561		0.0067	
12.2								0.0063									0.0005	
13	0.0040		0.1238	0.1435			0.3108	0.1034		0.0037	0.0011			0.1821	0.0313		0.0154	
13.2																	0.0015	
14	0.0006	0.0005	0.0517	0.0139			0.3942	0.0316		0.0990	0.0011			0.0140	0.0031		0.0376	
14.2																	0.0021	
15	0.0006	0.0497	0.0016	0.0006			0.1598	0.0028	0.0005	0.3141	0.0011			0.0014			0.0499	
15.2											0.0164						0.0036	
16	0.0058	0.0456	0.0005				0.0611		0.0387	0.2823	0.0055						0.0766	
16.1		0.0015																
16.2																	0.0005	
17	0.0727	0.0928					0.0044		0.2085	0.1859	0.0669						0.0787	
17.2																		
17.3		0.0035																
18	0.1136	0.2008					0.0005		0.0860	0.0996	0.0592						0.1034	
18.2											0.0044							
18.3		0.0051																
19	0.1834	0.1653							0.1216	0.0090	0.0987						0.0957	
19.2											0.0055						0.0010	
20	0.2630	0.1760							0.1307	0.0016	0.1228						0.0607	
20.2											0.0329						0.0021	
21	0.1130	0.0979							0.0575		0.1305						0.0165	
21.2											0.0154						0.0190	
22	0.0963	0.0720							0.0707		0.1732						0.0175	
22.2											0.0121						0.0180	
23	0.0634	0.0588							0.1180		0.1261					0.0005	0.0015	
23.2											0.0044						0.0283	
24	0.0392	0.0218							0.0905		0.0724					0.0205		
24.2											0.0011						0.0221	
25	0.0260	0.0086							0.0580							0.1951	0.0005	
25.2																	0.0298	
26	0.0029								0.0173							0.2395	0.0015	
26.2																	0.0509	
27	0.0006								0.0020							0.2373		
27.2																	0.0741	
28	0.0006										0.0395	0.0033				0.1578		
28.2											0.0011						0.0597	
29											0.0066	0.0123				0.1232		
29.2																	0.0484	
30											0.0022	0.0357				0.0249	0.0010	
30.2												0.0006						
30.3																	0.0314	
31												0.0848				0.0011		
31.2																	0.0237	
32												0.1172						
32.2																	0.0108	
33												0.1088					0.0005	
33.2																	0.0036	
34												0.1451						
34.2																		
35												0.1077						
35.2																	0.0021	
36												0.0949						
36.2																		
37												0.0887						
38												0.0837						
39												0.0497						
40												0.0290						
41												0.0246						
42												0.0106						
43												0.0017						
44												0.0011						
45												0.0006						

Supplementary Table 5 - Allele frequencies of eighteen autosomal STR in Northeast population of Brazil part b

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105						0.0005												
109						0.0161												
113						0.0348												
117						0.1589												
121						0.0862												
125						0.0909												
129						0.3287												
133						0.1454												
137						0.1096												
141						0.0260												
145						0.0031												
244					0.0130													
248					0.0206													
252					0.0584													
256					0.1548													
260					0.2554													
264					0.2251													
268					0.1872													
272					0.0595													
276					0.0184													
280					0.0032													
284					0.0043													
288																		
350													0.0016					
354													0.0016					
358													0.1176					
362													0.1036					
366													0.3134					
370													0.3220					
374													0.1084					
378													0.0248					
382													0.0038					
386													0.0027					
390													0.0005					
N	867	986	957	899	462	963	917	711	983	944	456	896	927	357	974	925	972	971
OH (%)	0.819	0.865	0.764	0.792	0.781	0.801	0.733	0.755	0.870	0.765	0.893	0.897	0.766	0.689	0.796	0.802	0.938	0.800
EH (%)	0.851	0.870	0.784	0.792	0.818	0.816	0.715	0.759	0.882	0.768	0.897	0.906	0.762	0.723	0.803	0.808	0.942	0.799
P	0.022	0.011	0.076	0.526	0.055	0.206	0.063	0.815	0.131	0.956	0.810	0.946	0.517	0.225	0.107	0.371	0.075	0.001
MP	0.040	0.031	0.080	0.077	0.058	0.055	0.134	0.090	0.026	0.091	0.021	0.017	0.094	0.121	0.071	0.066	0.008	0.074
Exp. as 1 in	24.829	32.035	12.512	13.061	17.263	18.096	7.490	11.101	38.836	11.030	46.811	57.915	10.605	8.289	14.181	15.217	122.317	13.458
PIC	0.835	0.856	0.754	0.760	0.792	0.795	0.667	0.724	0.870	0.731	0.887	0.897	0.725	0.676	0.774	0.779	0.936	0.767
PD	0.960	0.969	0.920	0.923	0.942	0.945	0.866	0.910	0.974	0.909	0.979	0.983	0.906	0.879	0.929	0.934	0.992	0.926
PE	0.635	0.725	0.534	0.584	0.565	0.600	0.481	0.519	0.734	0.535	0.780	0.790	0.534	0.412	0.591	0.603	0.867	0.599
TPI	2.761	3.707	2.117	2.404	2.287	2.508	1.871	2.043	3.840	2.126	4.653	4.870	2.136	1.608	2.447	2.527	7.700	2.503

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 6

Supplementary Table 6 - Allele frequencie of eighteen autosomal STR loci in Midwest population of Brazil part a

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																		
5																		
6								0.0106										0.2364
7																		0.2636
7.1																		
8			0.0877	0.0106				0.0106						0.0294	0.1842			0.1455
9			0.0702	0.1702				0.0638						0.0147	0.1667			0.1182
9.3																		0.2364
10			0.0351	0.1064				0.0426						0.0588	0.2193			
10.2																		
11			0.2544	0.2021				0.4468						0.3529	0.2807			
11.2																		
12			0.3246	0.2979			0.0566	0.2553						0.4118	0.1404			
12.2																		
13			0.1930	0.1596			0.3585	0.1277		0.0096				0.1324	0.0088			
13.2																		
14	0.0119		0.0351	0.0532			0.3491	0.0319		0.1250							0.0182	
14.2																		
15		0.0893					0.1792	0.0106		0.2692							0.0364	
15.2											0.0119							
16	0.0119	0.0625					0.0472		0.0526	0.3077	0.0119						0.0727	
16.1																		
16.2																		
17	0.0833	0.0893					0.0094		0.1930	0.1538	0.0476						0.0545	
17.2																		
17.3																		
18	0.1071	0.2054							0.1140	0.1250	0.0595						0.1000	
18.2																		
18.3																		
19	0.1667	0.0982							0.1579	0.0096	0.0952						0.0818	
19.2																		
20	0.2738	0.2232							0.1316		0.1071						0.0455	
20.2																	0.0091	
21	0.1310	0.0893							0.0789		0.2143						0.0455	
21.2											0.0119						0.0182	
22	0.0833	0.0714							0.0789		0.2262						0.0091	
22.2											0.0119						0.0273	
23	0.0714	0.0714							0.0877		0.1310						0.0091	
23.2											0.0119						0.0091	
24	0.0476								0.0526		0.0238					0.0357		
24.2																	0.0091	
25	0.0119								0.0351							0.1607		
25.2																	0.0182	
26									0.0175							0.2500		
26.2																	0.0455	
27																0.2321		
27.2																	0.1000	
28											0.0357					0.0714		
28.2																	0.1273	
29																0.1964		
29.2																	0.0727	
30												0.0521				0.0536		
30.2																	0.0455	
30.3																		
31												0.1042						
31.2																	0.0091	
32												0.0938						
32.2																	0.0182	
33												0.1250						
33.2																	0.0091	
34												0.1250						
34.2																		
35												0.0833						
35.2																	0.0091	
36												0.1562						
36.2																		
37												0.1042						
38												0.1042						
39												0.0417						
40																		
41												0.0104						
42																		
43																		
44																		
45																		

Supplementary Table 6 - Allele frequencie of eighteen autosomal STR loci in Midwest population of Brazil part b

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105																		
109						0.0182												
113						0.0273												
117						0.1273												
121						0.1091												
125						0.0727												
129						0.2909												
133						0.1636												
137						0.1545												
141						0.0182												
145						0.0182												
248					0.0238													
252					0.0357													
256					0.1429													
260					0.2619													
264					0.2143													
268					0.2143													
272					0.0714													
276					0.0238													
280					0.0119													
284																		
288																		
350																		
354																		
358													0.1442					
362													0.1250					
366													0.2500					
370													0.3365					
374													0.0962					
378													0.0385					
382																		
386													0.0096					
390																		
N	42	56	57	47	42	55	53	47	57	52	42	48	52	57	56	55	55	
OH (%)	0.857	0.839	0.737	0.766	0.690	0.891	0.679	0.702	0.895	0.846	0.785	0,938	0.654	0.706	0.789	0.804	0.945	0.836
EH (%)	0.857	0.868	0.785	0.810	0,821	0.837	0.719	0.719	0.888	0.785	0.867	0,901	0.784	0.694	0.799	0.817	0.940	0.791
P	0.906	0.767	0.897	0.111	0.192	0.939	0.818	0.777	0.141	0.640	0.472	0.599	0.193	0.775	0.800	0.534	0.025	0.774
MP	0.051	0.045	0,082	0.087	0.070	0.061	0.128	0.121	0.043	0.098	0.0488	0.040	0.087	0.151	0.081	0.075	0.030	0.091
Exp. as 1 in	19.600	22.400	12.169	11.446	14.226	16.351	7.825	8.273	23.043	10.165	20.512	25.043	11.556	6.644	12.354	13.288	33.362	10.921
PIC	0.830	0.845	0.745	0.774	0.785	0.809	0.661	0.673	0.869	0.7436	0.842	0.881	0.745	0.628	0.759	0.783	0.928	0.749
PD	0.949	0.955	0.918	0.913	0.930	0.939	0.872	0.879	0.957	0.902	0.951	0.960	0.913	0.849	0.919	0.925	0.970	0.908
PE	0.709	0.674	0.488	0.537	0.414	0.777	0.397	0.432	0.785	0.687	0.573	0.872	0.361	0.437	0.580	0.607	0.891	0.668
TPI	3.500	3.111	1.900	2.136	1.6154	4.583	1.559	1.679	4.750	3.250	2.333	8.000	1.444	1.700	2.375	2.545	9.333	3.056

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 7

Supplementary Table 7 - Allele frequencie of eighteen autosomal STR loci in South population of Brazil part a

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																		
5																		0.0029
6														0.0132				0.1954
7								0.0041						0.0132	0.0231			0.1810
7.1																		
8			0.1175	0.0189				0.0285						0.0088	0.1734		0.0029	0.1322
9			0.0783	0.1289			0.0034	0.0528						0.0219	0.0925			0.1638
9.3																		0.3075
10			0.0753	0.0723				0.0732						0.0614	0.2775			0.0172
10.2																		
11	0.0034		0.2831	0.2830			0.0034	0.2805						0.3465	0.2399			
11.2																		
12			0.3042	0.3208			0.0552	0.3659		0.0030				0.3465	0.1734		0.0058	
12.2																		
13	0.0034		0.0813	0.1447			0.2966	0.1545						0.1711	0.0202		0.0174	
13.2																		
14		0.0058	0.0602	0.0283			0.4448	0.0407		0.0823				0.0175			0.0349	
14.2																		
15		0.0289		0.0031			0.1276			0.2652							0.0494	
15.2											0.0038							
16		0.0491					0.0690		0.0514	0.3018	0.0038						0.1047	
16.1		0.0029																
16.2																		
17	0.0445	0.1069							0.2457	0.2104	0.0808						0.0930	
17.2																		
17.3		0.0116																
18	0.0616	0.1821							0.1200	0.1280	0.0962						0.0756	
18.2																		
18.3		0.0087																
19	0.1712	0.1532							0.1114	0.0091	0.1038						0.0640	
19.2																		
20	0.3596	0.1618							0.1143		0.1038						0.0581	
20.2											0.0077						0.0058	
21	0.1336	0.1098							0.0486		0.1423						0.0378	
21.2											0.0077						0.0145	
22	0.0822	0.0780							0.0457		0.1731						0.0087	
22.2																	0.0262	
23	0.0445	0.0434							0.1143		0.1538						0.0058	
23.2																	0.0262	
24	0.0479	0.0405							0.0343		0.0846					0.0231		
24.2																	0.0378	
25	0.0411	0.0173							0.0971							0.1676		
25.2																	0.0203	
26	0.0068								0.0143							0.2139	0.0029	
26.2																	0.0465	
27									0.0029							0.2370		
27.2																	0.0872	
28											0.0192					0.1329		
28.2											0.0038						0.0669	
29											0.0115	0.0217				0.1936		
29.2																	0.0349	
30											0.0038	0.0109				0.0289		
30.2																	0.0465	
30.3																		
31												0.0761				0.0029		
31.2																	0.0174	
32												0.1268						
32.2																	0.0058	
33												0.1486						
33.2																		
34												0.0906						
34.2																		
35												0.0942					0.0029	
35.2																		
36												0.1123						
36.2																		
37												0.0942						
38												0.1196						
39												0.0471						
40												0.0290						
41												0.0181						
42												0.0036						
43																		
44																		
45												0.0072						

Supplementary Table 7 - Allele frequencie of eighteen autosomal STR loci in South population of Brazil part b

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105						0.0030												
109						0.0119												
113						0.0208												
117						0.1250												
121						0.0506												
125						0.0982												
129						0.3333												
133						0.1696												
137						0.1518												
141						0.0327												
145						0.0030												
248					0.0129													
252					0.0474													
256					0.1897													
260					0.2414													
264					0.2457													
268					0.2026													
272					0.0560													
276					0.0043													
280																		
284																		
288																		
350																		
354																		
358													0.1166					
362													0.1012					
366													0.3037					
370													0.2791					
374													0.1503					
378													0.0399					
382													0.0061					
386													0.0031					
390																		
N	146	173	166	159	116	168	145	123	175	164	130	138	163	34	114	173	172	174
OH (%)	0.774	0.913	0.777	0.786	0.802	0.821	0.690	0.789	0.891	0.780	0.869	0.862	0.798	0.719	0.763	0.792	0.948	0.782
EH (%)	0.808	0.883	0.794	0.776	0.802	0.810	0.692	0.756	0.871	0.773	0.884	0.902	0.784	0.729	0.798	0.816	0.94298	0.792
P	0.802	0.183	0.660	0.847	0.887	0.564	0.105	0.472	0.890	0.851	0.086	0.440	0.859	0.554	0.825	0.654	0.576	0.660
Exp. as 1 in	16.971	31.015	12.757	11.383	13.483	15.207	6.208	8.781	29.139	10.916	29.754	42.,509	11.856	8.450	14.232	16.082	81.667	12.807
PIC	0.785	0.869	0.763	0.739	0.768	0.785	0.641	0.717	0.856	0.734	0.869	0.890	0.750	0,6800741	0.765	0.786	0.938	0.758
PD	0.941	0.968	0.922	0.912	0.926	0.934	0.839	0.886	0.966	0.908	0.966	0.976	0.916	0.882	0.930	0.938	0.988	0.922
PE	0.552	0.823	0.557	0.574	0.602	0.639	0.412	0.578	0.778	0.563	0.733	0.719	0.594	0.459	0.532	0.584	0.895	0.565
TPI	2.212	5.767	2.243	2.334	2.522	2.800	1.611	2.365	4.605	2.278	3.824	3.632	2.470	1.781	2.110	2.403	9.722	2.289

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

Supplementary Table 8

Supplementary Table 8 - Allele frequencie of eighteen autosomal STR loci in Southeast population of Brazil part a

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
4																		
5								0.0007							0.0005			0.0010
6								0.0088						0.0069				0.1898
7				0.0006				0.0022						0.0059	0.0132			0.2711
7.1														0.0020				
8			0.0789	0.0246			0.0005	0.0236						0.0168	0.1515			0.1538
9	0.0012		0.0805	0.1693				0.0811						0.0325	0.1170			0.1665
9.3																		0.2132
10	0.0018		0.0459	0.0818			0.0027	0.0966						0.0592	0.2893		0.0010	0.0046
10.2								0.0007										
11	0.0047		0.2972	0.3009			0.0016	0.3060		0.0017				0.3254	0.2310		0.0031	
11.2								0.0066									0.0005	
12	0.0047		0.3137	0.2460			0.0571	0.3127		0.0033				0.3491	0.1722		0.0072	
12.2								0.0052									0.0010	
13	0.0018		0.1357	0.1568			0.2999	0.1202	0.0005	0.0072	0.0008			0.1864	0.0228		0.0154	
13.2																	0.0005	
14	0.0012	0.0005	0.0464	0.0200			0.3938	0.0324	0.0005	0.0945	0.0016			0.0128	0.0025		0.0257	
14.2																	0.0010	
15	0.0035	0.0505	0.0015				0.1734	0.0029	0.0010	0.2928				0.0030			0.0565	
15.2											0.0256						0.0026	
16	0.0129	0.0409					0.0678		0.0435	0.2790	0.0085						0.0740	
16.1		0.0020																
16.2																	0.0005	
17	0.0959	0.0959					0.0027		0.2042	0.2193	0.0660						0.0843	
17.2											0.0008						0.0005	
17.3		0.0061																
18	0.1099	0.2119							0.0642	0.0917	0.0691						0.1043	
18.2											0.0070							
18.3		0.0055																
19	0.1684	0.1559					0.0005		0.1294	0.0099	0.1025						0.0940	
19.2											0.0062							
20	0.2450	0.1473							0.1304	0.0006	0.1180						0.0576	
20.2											0.0210						0.0036	
21	0.1205	0.0893							0.0723		0.1452						0.0226	
21.2											0.0272						0.0231	
22	0.0860	0.0964							0.0799		0.1615						0.0139	
22.2											0.0116						0.0154	
23	0.0643	0.0621							0.1132		0.1297					0.0005	0.0036	
23.2											0.0085						0.0190	
24	0.0409	0.0217							0.0809		0.0466					0.0159	0.0010	
24.2											0.0039						0.0242	
25	0.0316	0.0131							0.0597							0.1409		
25.2																	0.0303	
26	0.0047	0.0010							0.0182							0.2352		
26.2																	0.0488	
27									0.0020							0.2442		
27.2																	0.0704	
28	0.0012										0.0272	0.0024				0.1573	0.0010	
28.2																	0.0601	
29											0.0109	0.0142				0.1732	0.0005	
29.2																	0.0653	
30											0.0008	0.0385				0.0286		
30.2												0.0006						
30.3																	0.0339	
31												0.0729				0.0026	0.0005	
31.2												0.0012					0.0170	
32												0.1226				0.0005		
32.2																	0.0092	
33												0.1309						
33.2																	0.0031	
34												0.1286				0.0011		
34.2																	0.0005	
35												0.1167					0.0005	
35.2																	0.0015	
36												0.1037					0.0005	
36.2																	0.0005	
37												0.0776						
38												0.0895						
39												0.0474						
40												0.0243						
41												0.0160						
42												0.0089						
43												0.0030						
44												0.0012						
45																		

Supplementary Table 8 - Allele frequencie of eighteen autosomal STR loci in Southeast population of Brazil part b

Alelo	D10S1237	D12S391	D13S317	D16S539	D16S753	D21S1437	D22S534	D22S689	D2S1338	D3S1358	D3S2387	D3S2406	D5S2503	D5S818	D7S820	D9S938	SE33	THO1
105						0.0005												
109						0.0169												
113						0.0406												
117						0.1607												
121						0.0888												
125						0.1006												
129						0.3306												
133						0.1304												
137						0.1027												
141						0.0267												
145						0.0015												
244					0.0219													
248					0.0180													
252					0.0438													
256					0.1862													
260					0.2488													
264					0.2152													
268					0.1792													
272					0.0532													
276					0.0203													
280					0.0094													
284					0.0031													
288					0.0008													
350													0.0011					
354													0.0040					
358													0.0909					
362													0.1194					
366													0.2989					
370													0.3331					
374													0.1183					
378													0.0269					
382													0.0040					
386													0.0023					
390													0.0011					
N	855	991	969	874	639	974	937	678	989	905	644	844	875	507	987	944	973	985
OH (%)	0.850	0.,864	0.760	0.799	0.812	0.817	0.710	0.795	0.875	0.751	0.882	0.871	0.746	0.746	0.787	0.780	0.939	0.759
EH (%)	0.862	0.874	0.778	0.788	0.820	0.817	0.717	0.777	0.884	0.771	0.898	0.903	0.763	0.733	0.796	0,810	0.941	0.794
P	0.408	0.275	0.790	0.079	0.083	0.228	0.324	0.129	0.078	0.579	0.782	0.098	0.801	0.177	0.991	0.194	0.051	0.001
MP	0.034	0.028	0.077	0.078	0.060	0.055	0.129	0.085	0.025	0.088	0.020	0.018	0.091	0.126	0.072	0.064	0.008	0.074
Exp. as 1 in	29.348	35.179	12.961	12.801	16.645	18.090	7.759	11.807	39.334	11.325	50.430	55.669	10.942	7.965	13.877	15.717	119.915	13.429
PIC	0.847	0.861	0.746	0.756	0.794	0.796	0.670	0.745	0.872	0.734	0.888	0.894	0.727	0.688	0.766	0.782	0.934	0.761
PD	0.966	0.972	0.923	0.922	0.940	0.945	0.871	0.915	0.975	0.912	0.980	0.982	0.909	0.874	0.928	0.936	0.992	0.926
PE	0.695	0.722	0.526	0.596	0.622	0.631	0.443	0.590	0.744	0.512	0.759	0.736	0.503	0.502	0.576	0.562	0.869	0.526
TPI	3.340	3.670	2.079	2.483	2.663	2.736	1.722	2.439	3.988	2.011	4.237	3.872	1.970	1.965	2.350	2.269	7.788	2.078

N, number of individuals per loci; OH, observed heterozygosity; EH, expected heterozygosity; P , P value (0.00034) after Bonferroni correction; MP, matching probability; PD, power of discrimination; PIC, polymorphism information content; PE, probability exclusion; TPI, typical paternity index.

4 - DISCUSSÃO

Durante décadas as análises forenses baseavam-se em sistemas multiplex compostos por 10-15 STRs que forneciam informações genéticas suficientes para elucidar casos simples de verificação de parentesco. Estes STRs eram à base dos bancos de dados europeus e norte-americano como o CODIS. No entanto, recentemente, outros conjuntos de STRs autossômicos vêm ganhando destaque para aumentar as chances de resolução de casos complexos de verificação de parentesco. Todas essas aplicações requerem tanto sensibilidade forense quanto um número maior de marcadores genéticos disponíveis para obter probabilidades suficientemente informativas (Phillips *et al*., 2014; Asamura *et al.*, 2007). Na presente tese, caracterizamos nove novos STRs e analisamos dois novos conjuntos de marcadores, totalizando dezoito marcadores, que são mesclados com os marcadores do sistema CODIS com o objetivo de caracterizar e avaliar a informatividade desses conjuntos para análises forenses e estudos populacionais.

Conhecer a localização cromossômica, o motivo de repetição, os alelos disponíveis e o tamanho do produto de PCR são imprescindíveis para a padronização de novos sistemas multiplex (Buttler, 2007). Na primeira etapa do nosso estudo foram caracterizados molecularmente nove novos marcadores STRs, todos os marcadores possuem quatro bases em cada motivo de repetição e são classificados como tetranucleotídeos. Em aplicações forenses, é mais comum utilizar STRs tetranucleotídeos, pois estes apresentam menor número de problemas com picos *stutter* (Jobling & Gill, 2004). Os picos *stutter* são artefatos resultantes da amplificação de STRs, caracterizados pela presença de uma unidade de repetição mais curta em relação ao alelo principal (Buttler, 2007; Seo *et al.*, 2014). Apenas dois STRs possuem motivo de repetição imperfeito, mais de um tipo de motivo repetição para o mesmo STR, este tipo de motivo de repetição é formado por mutações pontuais e pequenas inserções e deleções durante a evolução de cada loco (Pemberton *et al.*, 2009).

As evidências do DNA em análises forenses e teste de paternidade são baseados nas interpretações de similaridades e diferenças em cada marcador genético. Nos testes de paternidade, as diferenças nos marcadores entre o suposto pai e o filho definem a exclusão da paternidade. No entanto, mutações espontâneas na linhagem germinativa do suposto pai para um determinado marcador são naturais e promovem a alta variabilidade destes marcadores (Kayser & Sajantila, 2001). Os STRs possuem uma taxa de mutação média de 1,2 x 10^{-3} (Brinkmann *et al.*, 1998), no nosso estudo podemos observar uma variação na taxa de mutação entre 1 x 10^{-4} para o D9S938 e 2,2 x 10^{-3} Para o SE33. Os marcadores que possuem as taxas mais altas de mutação são aqueles com motivo de repetição imperfeito e possuem mais de 10 motivos de repetição, como SE33 (Wenda *et al.,* 2005), D12S391 (Lareu *et al*., 1996) e D3S2406. Os marcadores com sequências imperfeitas e com um maior número de repetições são mais susceptíveis a eventos mutacionais (Brinkamnn *et al.*, 1998; Pemberton *et al.*, 2009).

As análises estatísticas dos parâmetros forenses são usadas para auxiliar na interpretação de resultados de identificação genética e verificação de parentesco. Estas análises atribuem valor aos resultados obtidos e facilitam a resolução dos casos forenses (Huston, 1998). Os resultados obtidos para os conjuntos de marcadores aqui caracterizados com o objetivo de auxiliar nas análises com os kits convencionais apresentaram bons resultados comparados aos outros kits, como HDPlex e Powerplex ESX 17, desenvolvidos com o mesmo objetivo. Por exemplo, a probabilidade de correspondência (*Random match probability*) observada no nosso estudo foi de 4,036 x 10^{-24} enquanto no HDPlex (Qiagen®), composto por 13 locos, variou entre 1,0 x 10^{-10} a 3,3 x 10^{-14} (Phillips *et al.*, 2014). Quando comparamos com o Powerplex ESX 17 (Promega®), composto por 17 locos, confirmamos os bons resultados, a média da informação polimórfica contida nos marcadores do Powerplex ESX 17 foi de 81,3% enquanto nos marcadores aqui caracterizados foi de 79,5% (Sousa *et al.*, 2014).

Bancos de dados populacionais são criados para manter a informação genética de cada indivíduo para um dado marcador. Estes bancos de dados são definidos por grupos étnicos e regiões geográficas porque os alelos podem ter diferentes frequências em diferentes populações (Huston, 1998). Sendo assim, a última etapa para validação dos conjuntos de marcadores é o estudo populacional baseado nestes marcadores que está descrito no capítulo II.

A estrutura genética é moldada ao longo do tempo pela interação de diversos fatores como seleção natural, deriva genética, mutação, migração, endogamia, efeito fundador, entre outros. Uma das formas de se avaliar a estrutura genética das populações é o uso dos índices de fixação (estatística F de Wright) como medidas de distâncias genéticas (Silva, 2010).

Na análise do F_{st} loco por loco (Tabela 1 – Capítulo II) todos os marcadores apresentaram valores baixos e apenas dois marcadores D9S938 e TH01 apresentaram valores significativos. Os resultados do F_{st} entre as populações analisadas par a par (Tabela 2 – Capítulo II) também apresentaram baixos valores entre as populações brasileiras, sendo que dos 10 valores computados 4 apresentaram valores significativos. O valor mais alto foi observado entre a população Norte e Sul, o que condiz com a distância genética entre as duas populações. Os valores baixos entre as populações brasileiras foram relatados previamente (Lins, 2007).

Os outros dois componentes das estatísticas F são o F_{it} e F_{is} (Tabela 1 – Capítulo II) não apresentam valores significativos e assim confirmam os resultados observados nas análises do equilíbrio de Hardy-Weinberg, onde foi observado que após correção de Bonferroni todas as populações estão em equilíbrio.

Para finalizar as análises populacionais baseadas nos conjuntos de marcadores caracterizados nesse estudo verificamos a contribuição genética das populações africanas e europeias nas populações brasileiras. Para isso analisamos as populações do repositório Coriell para o painel de marcadores caracterizados. Foram analisadas populações europeia, africana e hispânica. As populações europeias e africanas foram estudas por que historicamente a população Brasileira é um produto do complexo processo de miscigenação que tem entre suas raízes principais estas populações. Infelizmente não tivemos acesso a populações ameríndias que completariam as raízes principais da formação da população brasileira. A população hispânica foi incluída por ser composta por indivíduos miscigenados.

Os resultados do F_{st} par a par (tabela 2 – Capítulo II) mostraram valores significativos para todas as comparações em relação à população Africana, sendo que a maior diferença foi em relação à população Sul. A população Europeia só não demonstrou valores significativos na comparação com as populações Centro-Oeste e Sul. Estes resultados foram confirmados analisando os resultados do STRUCTURE (Figura 2 – Capítulo II) onde verificamos que a contribuição europeia é maior que a africana variando entre 78,2% na população Nordeste e 82,6% na população Sul. Os resultados das populações brasileiras foram próximos ao resultado observado para os hispânicos miscigenados com 78,1% de contribuição europeia. Os resultados encontrados condizem com outros estudos com populações brasileiras onde foi observada uma contribuição europeia que varia entre 68% - Norte e 81% - Sul (Lins *et al*., 2010).

Ao final podemos concluir que os marcadores caracterizados são bons marcadores para elucidar casos forenses por se mostraram tão informativos quanto os marcadores do sistema CODIS. As análises destes novos marcadores auxiliarão na resolução de casos complexos de verificação de parentesco e casos *post-mortem*.

Nas análises populacionais foi possível verificar diferenças genéticas significativas entre as populações brasileiras. Ainda nas análises populacionais foi possível confirmar que a contribuição genética europeia foi maior que a africana durante o processo de formação da população brasileira.

5 - CONSIDERAÇÕES FINAIS

Na última década as análises forenses no Brasil tiveram um grande avanço. Os testes de DNA foram considerados como um dos eventos que mudaram a vida dos brasileiros, pois ela foi amplamente difundida junto ao direito de família. A junção dessa divulgação aliada aos avanços das tecnologias e diminuição dos custos com os testes fez com que os testes de verificação de parentesco se popularizassem, muitos casos são custeados pelo governo e outros órgãos públicos. Com essa popularização o número de casos complexos, como irmandades e vínculos genéticos familiares, aumentou significativamente. Sendo assim, os peritos responsáveis pela elucidação destes casos contam cada vez mais com os avanços tecnológicos e com estudos como este que além de validarem novos marcadores, trazem informações sobre a população brasileira.

6 - REFERÊNCIAS BIBLIOGRÁFICAS

Aguiar, S. M., et al. (2011). Rede Integrada de Bancos de Perfis Genéticos e a implantação do CODIS no Brasil. Congresso Brasileiro de Genética Forense, Porto Alegre.

Aguiar, V. R. C., et al. (2012). "Updated Brazilian STR allele frequency data using over 100,000 individuals: an analysis of CSF1PO, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA loci." Forensic Sci. Int. Genet **6**: 504–509.

Asamura, H., et al. (2007). "MiniSTR multiplex systems based on non-CODIS loci for analysis of degraded DNA samples." Forensic Science International **173**: 7–15.

Balding, D. J. (2005). Weight-of-evidence for forensic DNA profiles, Wiley and Sons Ltda, Chichester, UK. 185p.

Bossart, J. L. and Prowell D. P. (1998). "Genetic estimates of population structure and gene flow: limitations, lessons and new directions." Tree **13**: 202-206.

Brinkmann, B., et al. (1998). "Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat." Am. J. Hum. Genet. **62**: 1408–1415.

Butler, J. (2007). "Short tandem repeat typing technologies used in human identity testing." BioTechniques **43**.

Cabrero, C., et al. (1995). "Allele frequency distribution of four PCR-amplified loco in Spanish population." Forensic Science International **71**: 153-164.

Ellegren, H. (2004). "Microsatellites: simple sequences with complex evolution." Nature Reviews: Genetics **5**: 435-445.

Excoffier, L., et al. (1992). "Analysis of molecular variance inferred from metric distances among DNA haplotype: application of human mitochondrial DNA restriction data." Genetics **131**: 479-491.

Garofano, L. P., M., et al. (1999). "Italian population data on two new short tandem repeat loci: D2S1338 and Penta E." Forensic Sci Int. **105(2)**: 131-136.

Gjertson, D. W., et al. (2007). " ISFG: Recommendations on biostatistics in paternity testing." Forensic Sci Int Genet **3-4**: 223-231.

Hey, J. and Machado C. A. (2003). "The study of structured population – New hope for a difficult and divided sciene." Nature Reviews: Genetics **4**: 535-543.

Holsinger, K. E. and Weir B. S. (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}." Nature Reviews: Genetics **10**: 639-650.

Huston, K. A. (1998). "Statistical analysis of STR data." Profiles in DNA (Promega Corporation): 14-15.

Jobling, M. A. and Gill P. (2004). "Encoded evidence: DNA in forensic analysis." Nature Reviews: Genetics **5**: 739-752.

Kayser, M. and Sajantila A. (2001). "Mutation at Y-STR loci: implication for paternity testing and forensic analysis." Forensic Science International **118**: 116-121.

Lareu, M. V., et al. (1996). "A highly variable STR at the D12S391 locus." Int J Legal Med. **109(3)**: 134-138.

Lins, T. C., et al. (2010). "Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs." American Journal of Human Biology **22**: 187-192.

Lins, T. C. L. (2007). Impacto da miscigenação na aplicação do HapMap para a população brasileira avaliados nos genes PTPN22 e VDR. Brasília, Universidade Católica de Brasília. **Mestrado**.

Michalakis, Y. and Excoffier L. (1996). "A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci." Genetics **142**: 1061-1064.

Pemberton, T. J., et al. (2009). "Sequence determinants of human microsatellite variability." BMC Genomics **10**: 2-19.

Phillips, C., et al. (2014). "Global population variability in Qiagen Investigator HDplex STRs." Forensic Science International: Genetics **8**: 36-43.

Pritchard, J. K., et al. (2000). " Inference of population structure using multilocus genotype data." Genetics **155**(945-959).

Ridley, M. (2006). Evolução. Porto Alegre-RS, Editora Artmed. 752p.

Scliar, M. O., et al. (2012). "The population genetics of Quechuas, the largest native south american group: autosomal sequences, SNPs, and microsatellites evidence high level of diversity." Am J Phys Anthropol(147(3)): 443-451.

Seo, S. B., et al. (2014). "Reduction of stutter ratios in short tandem repeat loci typing of low copy number DNA samples." Forensic Science International: Genetics **8**: 213-218.

Silva, M. C. F. (2010). Padrões geográficos de ancestralidade genômica em Minas Gerais: o caso da doença falciforme. Departamento de Biologia Geral. Belo Horizonte, Universidade Federal de Minas Gerais. **Doutorado**.

Sousa, M. L. A. P. O., et al. (2014). "Population data of 16 autosomal STR loci of the Powerplex ESX 17 System in a Brazilian Population from the State of São Paulo." Forensic Science International: Genetics **11**: e15-e17.

Sun, H., et al. (2014). "Comparison of southern Chinese Han and Brazilian Caucasian mutation rates at autosomal short tandem repeat loci used in human forensic genetics." Int J Legal Med. **128(1)**: 1-9.

Weir, B. S., et al. (2006). "Genetic Relatedness Analysis: modern data and new challenges." Nature Reviews: Genetics **7**: 771-780.

Wenda, S., et al. (2005). "ACTBP2 (alias ACTBP8) is localized on chromosome 6 (band 6q14)." Forensic Sci. Int. **148**: 207-209.

Wright, S. (1951). "The genetical structure of populations." Ann. Eugen. **15**: 323-354.

7 - SITES ACESSADOS

FBI. Frequently Asked Questions (FAQs) on the CODIS Program and the National DNA Index System. The FBI Federal Bureau Investigation. 2013. Disponível em: <http://www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet> Acessado em 18/06/2013 às 21:40hrs.

Filho F A. A ciência dá a pista. Terra. Disponível em: <http://www.terra.com.br/istoe-temp/edicoes/2064/imprime140068.htm> Acessado em 05/09/2014 às 16:30hrs.

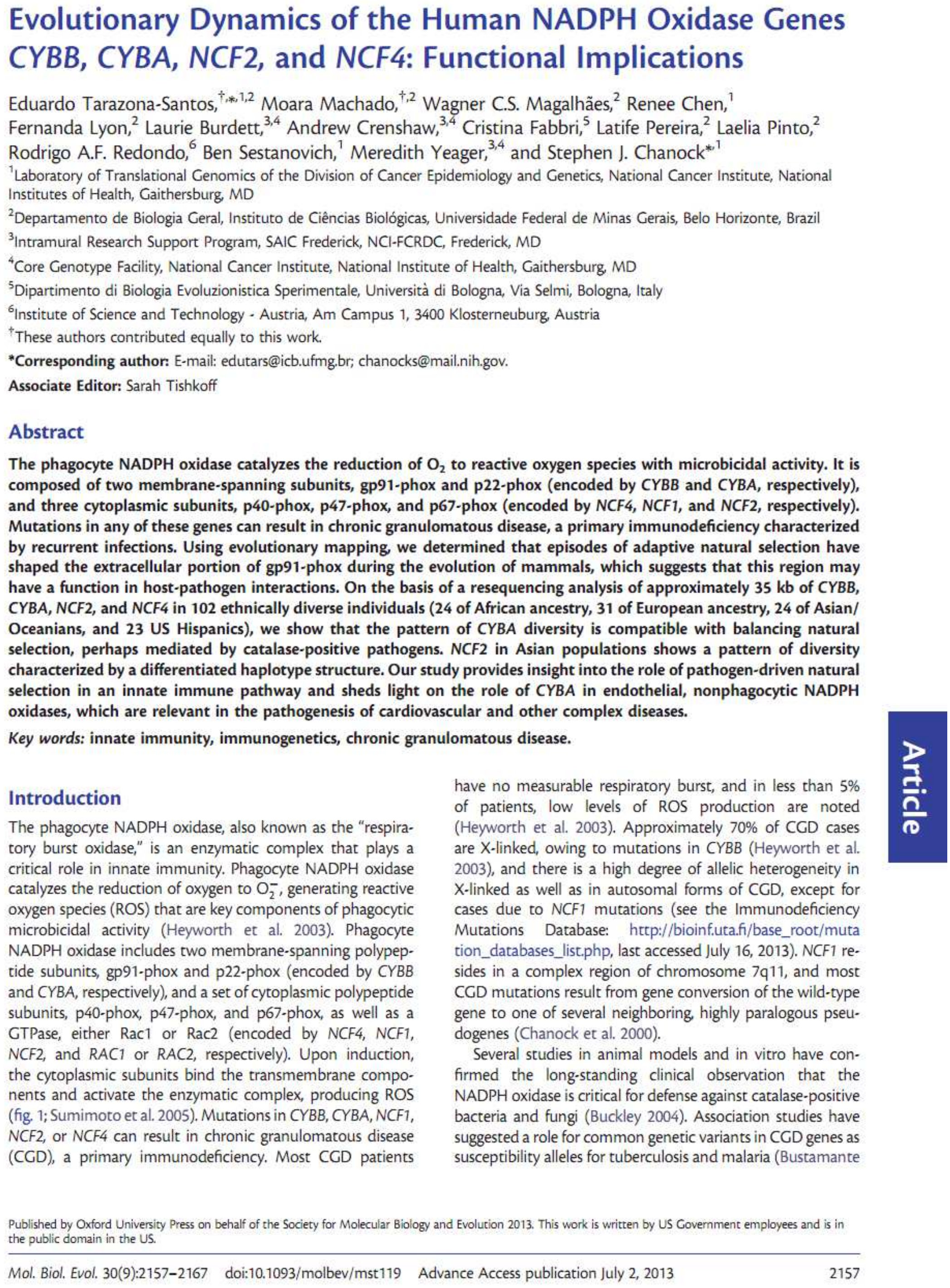
IBDFAM. Regulamentação do exame de DNA é tema de cartilha lançada pelo MPE-MG. IBDFAM Instituto Brasileiro de Direito de Família. 2009. Disponível em: <https://www.ibdfam.org.br/noticias/namidia/2856/Regulamenta%C3%A7%C3%A3o+do+exame+de+DNA+%C3%A9+tema+de+cartilha+lan%C3%A7ada+pelo+MPE-MG> Acessado em 05/09/2014 às 17:00hrs.

IBGE. Brasil 500 anos. Disponível em: <http://brasil500anos.ibge.gov.br/territorio-brasileiro-e-povoamento> Acessado em 10/11/2014 às 19:30hrs.

8 - ANEXOS

8.1 – OUTROS ESTUDOS

8.1.1 - *Evolutionary Dynamics of the Human NADPH Oxidase Genes CYBB, CYBA, NCF2, and NCF4: Functional Implications*



Evolutionary Dynamics of the Human NADPH Oxidase Genes *CYBB, CYBA, NCF2,* and *NCF4*: Functional Implications

Eduardo Tarazona-Santos,[†,*,1,2] Moara Machado,[†,2] Wagner C.S. Magalhães,[2] Renee Chen,[1] Fernanda Lyon,[2] Laurie Burdett,[3,4] Andrew Crenshaw,[3,4] Cristina Fabbri,[5] Latife Pereira,[2] Laelia Pinto,[2] Rodrigo A.F. Redondo,[6] Ben Sestanovich,[1] Meredith Yeager,[3,4] and Stephen J. Chanock*[1]

[1]Laboratory of Translational Genomics of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Gaithersburg, MD

[2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[3]Intramural Research Support Program, SAIC Frederick, NCI-FCRDC, Frederick, MD

[4]Core Genotype Facility, National Cancer Institute, National Institute of Health, Gaithersburg, MD

[5]Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Via Selmi, Bologna, Italy

[6]Institute of Science and Technology - Austria, Am Campus 1, 3400 Klosterneuburg, Austria

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: edutars@icb.ufmg.br; chanocks@mail.nih.gov.

Associate Editor: Sarah Tishkoff

Abstract

The phagocyte NADPH oxidase catalyzes the reduction of O_2 to reactive oxygen species with microbicidal activity. It is composed of two membrane-spanning subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*, respectively), and three cytoplasmic subunits, p40-phox, p47-phox, and p67-phox (encoded by *NCF4*, *NCF1*, and *NCF2*, respectively). Mutations in any of these genes can result in chronic granulomatous disease, a primary immunodeficiency characterized by recurrent infections. Using evolutionary mapping, we determined that episodes of adaptive natural selection have shaped the extracellular portion of gp91-phox during the evolution of mammals, which suggests that this region may have a function in host-pathogen interactions. On the basis of a resequencing analysis of approximately 35 kb of *CYBB*, *CYBA*, *NCF2*, and *NCF4* in 102 ethnically diverse individuals (24 of African ancestry, 31 of European ancestry, 24 of Asian/Oceanians, and 23 US Hispanics), we show that the pattern of *CYBA* diversity is compatible with balancing natural selection, perhaps mediated by catalase-positive pathogens. *NCF2* in Asian populations shows a pattern of diversity characterized by a differentiated haplotype structure. Our study provides insight into the role of pathogen-driven natural selection in an innate immune pathway and sheds light on the role of *CYBA* in endothelial, nonphagocytic NADPH oxidases, which are relevant in the pathogenesis of cardiovascular and other complex diseases.

Key words: **innate immunity, immunogenetics, chronic granulomatous disease.**

Article

Introduction

The phagocyte NADPH oxidase, also known as the "respiratory burst oxidase," is an enzymatic complex that plays a critical role in innate immunity. Phagocyte NADPH oxidase catalyzes the reduction of oxygen to O_2^-, generating reactive oxygen species (ROS) that are key components of phagocytic microbicidal activity (Heyworth et al. 2003). Phagocyte NADPH oxidase includes two membrane-spanning polypeptide subunits, gp91-phox and p22-phox (encoded by *CYBB* and *CYBA*, respectively), and a set of cytoplasmic polypeptide subunits, p40-phox, p47-phox, and p67-phox, as well as a GTPase, either Rac1 or Rac2 (encoded by *NCF4*, *NCF1*, *NCF2*, and *RAC1* or *RAC2*, respectively). Upon induction, the cytoplasmic subunits bind the transmembrane components and activate the enzymatic complex, producing ROS (fig. 1; Sumimoto et al. 2005). Mutations in *CYBB*, *CYBA*, *NCF1*, *NCF2*, or *NCF4* can result in chronic granulomatous disease (CGD), a primary immunodeficiency. Most CGD patients have no measurable respiratory burst, and in less than 5% of patients, low levels of ROS production are noted (Heyworth et al. 2003). Approximately 70% of CGD cases are X-linked, owing to mutations in *CYBB* (Heyworth et al. 2003), and there is a high degree of allelic heterogeneity in X-linked as well as in autosomal forms of CGD, except for cases due to *NCF1* mutations (see the Immunodeficiency Mutations Database: http://bioinf.uta.fi/base_root/mutation_databases_list.php, last accessed July 16, 2013). *NCF1* resides in a complex region of chromosome 7q11, and most CGD mutations result from gene conversion of the wild-type gene to one of several neighboring, highly paralogous pseudogenes (Chanock et al. 2000).

Several studies in animal models and in vitro have confirmed the long-standing clinical observation that the NADPH oxidase is critical for defense against catalase-positive bacteria and fungi (Buckley 2004). Association studies have suggested a role for common genetic variants in CGD genes as susceptibility alleles for tuberculosis and malaria (Bustamante

Mol. Biol. Evol. 30(9):2157–2167 doi:10.1093/molbev/mst119 Advance Access publication July 2, 2013 2157

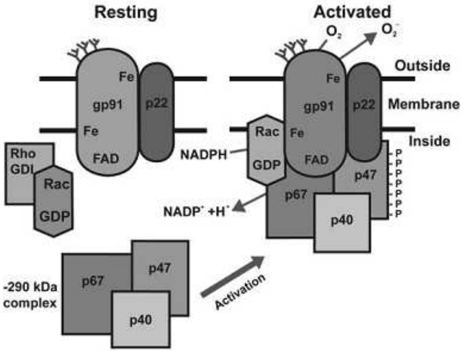


Fig. 1. Components of the phagocyte NADPH oxidase. Representation of the inactivated (left) and activated (right) forms of the phagocyte NADPH oxidase components, reproduced from Heyworth et al. (2003). The activated form is responsible for the respiratory burst. The proteins (and genes) are gp91 (*CYBB*, Xp21.1), p22 (*CYBA*, 16q24), p67 (*NCF2*, 1q25), p40 (*NCF4*, 22q13.1), and p47 (*NCF1*, 7q11.23).

et al. 2011), as well as for immune related diseases such as Crohn's disease and lupus, as identified in genome-wide association studies (GWAS) in European populations (Rioux et al. 2007; Roberts et al. 2008; Jacob et al. 2012). Besides the phagocyte NADPH oxidase, other NADPH oxidases with different functions are expressed in a variety of nonphagocytic cells, including the endothelium, and have been implicated in cardiovascular and renal disease. Although p22-phox (encoded by *CYBA*) is a protein component shared by several of these NADPH oxidases (also called Nox), other more specific protein subunits are encoded by different Nox genes homologous to the genes coding for the phagocytic subunits (Sumimoto et al. 2005; San José et al. 2008). Although these nonphagocytic NADPH oxidases normally produce less O_2^-, even small imbalances in ROS levels may cause tissue damage due to oxidative stress, which is correlated with the pathogenesis of gout, chronic obstructive pulmonary disease, rheumatoid arthritis, and cardiovascular diseases (Brandes and Kreuzer 2005). Therefore, variants in NADPH oxidase genes may have pleiotropic effects across a spectrum of disorders (Santiago et al. 2012).

Despite the involvement of the NADPH oxidase in a range of clinically relevant phenotypes, our knowledge of the sequence diversity of NADPH genes mostly derives from CGD patients. Although targeted SNP genotyping has been performed in the context of association studies for *CYBA* (Bedard et al. 2009) and *NCF4* (Olsson et al. 2007), none of the large-scale resequencing efforts, such as Seattle SNPs (http://pga.gs.washington.edu/, last accessed July 16, 2013), Innate Immunity PGA (http://www.pharmgat.org/IIPGA2/index_html, last accessed July 16, 2013), and the Cornell–Celera initiative (Bustamante et al. 2005), have included the NADPH oxidase genes, and the coverage of these genes for the current release of the 1000 Genomes Project remains low for most of the studied individuals (1000 Genomes Project Consortium et al. 2012; average coverage and their standard deviations on May 2013 are *CYBB*: 4.0 ± 2.2, *CYBA*: 3.5 ± 2.0, *NCF2*: 4.9 ± 2.7, and *NCF4*: 4.9 ± 2.7). Although GWAS have identified common variants that contribute to complex phenotypes, a component of missing heritability of common diseases due to rare variants that are detectable only by resequencing is emerging. In this study, we analyzed the pattern of sequence diversity of four of the NADPH genes (*CYBB*, *CYBA*, *NCF2*, and *NCF4*) between mammalian species and in human populations by resequencing these genes in 102 ethnically diverse individuals. We interpreted our results in terms of evolutionary histories, by addressing the action of natural selection and focusing on two temporal scales: mammalian evolution and recent human evolution. We excluded *NCF1* from our study because its high homology with its pseudogenes prevents reliable sequencing in individual samples (Chanock et al. 2000). Several studies have shown the importance of natural selection on the evolution of immunity genes at both the interspecific (Kosiol et al. 2008) and population levels (Ferrer-Admetlla et al. 2008; Barreiro et al. 2009; Barreiro and Quintana-Murci 2010). By definition, variants under natural selection are associated with different reproductive efficiencies (fitness) of their carriers and contribute to phenotype variability; therefore, they may be biomedically relevant by influencing the susceptibility to rare or common diseases. The goals of this study are as follows: 1) to determine whether the pattern of diversity of human phagocyte NADPH genes reflects the action of different types of natural selection, 2) to elucidate the evolutionary dynamics of NADPH genes at the temporal scales of mammals and humans, and 3) to understand the biomedical implications of this evolutionary process in human populations.

Results

Molecular Evolution of NADPH Genes along Mammalian Phylogeny

We examined signatures of natural selection across the coding regions of NADPH genes by analyzing sequences from the complete genomes of 29 mammals listed in the Entrez and Ensembl databases (Lindblad-Toh et al. 2011, one sequence for each species, see supplementary material, Supplementary Material online for details) and comparing the amount of nonsynonymous and synonymous substitutions (Nielsen et al. 2005). When comparing a set of homologous sequences from different species, most of the observed differences are *fixed*; that is, the differences are monomorphic within a species because enough time has passed for the observed variant to appear, increase its frequency and reach a frequency of 1 (Kimura 1974). We compared the number of fixed synonymous substitutions (dS, assumed to be neutral) and fixed nonsynonymous substitutions (dN, for which we test the hypothesis of natural selection) between species using the parameter $\omega = dN/dS$, which is informative of the action of natural selection at the inter-specific level (Yang 2007a). Under neutral evolution of nonsynonymous substitutions, these substitutions fix at the same rate as synonymous substitutions, and therefore $dN \approx dS$ and $\omega \approx 1$. If nonsynonymous substitutions tend to be deleterious, purifying selection maintains the substitutions at low frequencies and prevents fixation at the same rate as synonymous substitutions, resulting in $dN < dS$ and $\omega < 1$. On the other hand, if episodes of positive natural selection (that raise the frequency of beneficial variants) are frequent, nonsynonymous substitutions increase in frequency and fix more rapidly than neutral synonymous substitutions, thus, $dN > dS$ and $\omega > 1$. We used the maximum likelihood framework developed by Yang (2007a) to estimate ω for the NADPH oxidase genes under a variety of evolutionary models, as implemented in the PAML software (Yang 2007b). This approach allows inferences about the evolution of a coding region along an interspecific phylogeny and maps the codons that have evolved under strong/weak purifying selection, neutrality, or adaptive positive selection (see supplementary material, Supplementary Material online for details).

In general, PAML evolutionary models that allow a combination of purifying selection and neutrality are reasonably realistic. These models are nested with respect to models that also incorporate positive selection at the cost of adding new parameters. We evaluated the improvements in the goodness of fit of the data using the latter model with respect to the former models by applying a likelihood ratio test (LRT). After fitting the data to the most appropriate evolutionary model, Naive Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB) approaches were used to infer the ω parameter for each codon.

For the 29 species of mammals considered, we filtered based on quality control (supplementary table S1, Supplementary Material online) and analyzed 570 codons of *CYBB* in 26 species, 198 codons of *CYBA* in 16 species, 526 codons of *NCF2* in 23 species, and 339 codons of *NCF4* in 20 species (see supplementary material, Supplementary Material online for further details including the species and sequences used for the analyses, the parameter estimations for the different models and the LRT results). Here, we only present the results of model M3 of Yang (2007a) with three ($K = 3$) classes of ω (fig. 2). This model allows for different ω classes, including the possibility of positive selection, and is a reasonable way of presenting the results for the four genes under the same model. Moreover, in all cases, the data fit better with model M3 than the nested and simpler M0 or M2 models presented by Yang (2007a).

The results of this analysis for *CYBB*, *CYBA*, *NCF2*, and *NCF4* are presented in figure 2, which shows the type of natural selection (i.e., based on the estimated ω) for each codon that most likely predominated during mammalian evolution. For this temporal scale, *CYBA*, *NCF2*, and *NCF4* coding regions have evolved driven by a combination of different levels of purifying natural selection. Overall, the average and standard deviation values for these genes are $\omega_{NCF2} = 0.256 \pm 0.227$, $\omega_{NCF4} = 0.126 \pm 0.140$, and $\omega_{CYBA} = 0.109 \pm 0.116$.

Our most striking result is for *CYBB*, which presents a wide spectrum of mutations that account for >70% of CGD patients. Although we would predict that purifying selection on genes involved in Mendelian diseases (Blekhman et al. 2008) would yield similar results for *CYBB* and other NADPH components, we observed a different pattern. In general, *CYBB* is a conserved gene, but 6% of its codons show evidence of positive natural selection (supplementary table S2, Supplementary Material online; fig. 2). In a genome-wide survey performed by Kosiol et al. (2008), they have reported *CYBB* as a gene showing a signal of positive natural selection. More importantly, by evolutionary mapping, we show here for the first time that most of these positive selection events map to the small extracellular portion of this protein (fig. 3). The proximity of these inferred episodes of positive natural selection to glycosylation sites in gp91 is noteworthy considering the importance of the glycome in immunity (Marth and Grewal 2008).

Population Genetics of NADPH Genes

We sequenced *CYBB*, *CYBA*, *NCF2*, and *NCF4* in a publicly available panel that includes 24 individuals of African ancestry, 31 Europeans, 24 Asian/Oceanians, and 23 admixed Latin Americans (i.e., Hispanics). This panel is a suboptimal representation of the worldwide population, a limitation that is common to most human genomic diversity projects focused on SNP genotyping or resequencing efforts. However, based on how human genetic diversity is apportioned within (>85%) and between (<15%) populations (Lewontin 1972; 1000 Genomes Project Consortium 2010), even studies using suboptimal sampling are informative about the genetic structure of human populations and serve to critically identify the role of evolutionary factors in human genetic diversity (Kimura 1974; Nielsen et al. 2005; 1000 Genomes Project Consortium 2010). All the raw results are available as supplementary material, Supplementary Material online, and at the *SNP500Cancer* project homepage (http://variantgps.nci.nih.gov/cgfseq/pages/snp500.do, last accessed July 16, 2013) or can be downloaded from the DIVERGENOME platform (Magalhães et al. 2012, http://www.pggenetica.icb.ufmg.br/divergenome/, last accessed July 16, 2013).

To ascertain which combination of evolutionary factors has shaped the diversity of NADPH genes, we assessed the pattern of nonsynonymous and synonymous polymorphisms, as well as intra- and interpopulation diversity for NADPH genes, and tested the null hypothesis of neutrality: that patterns of diversity may be explained by considering only the demographic history of human populations and the mutation and recombination rates of each locus.

Nonsynonymous polymorphisms are underrepresented in the human genome and usually occur at low frequencies when present, reflecting the action of purifying natural selection (1000 Genomes Project Consortium 2010). By resequencing, Tarazona-Santos et al. (2008) did not observe common nonsynonymous polymorphisms for *CYBB*. This result is consistent with purifying natural selection acting on X-chromosome genes due to the exposure of deleterious recessive mutations to natural selection in hemizygous males. Thus, substitutions in the coding region of *CYBB* should be rare in human populations and are seldom captured by studies with small sample sizes. Interestingly, the lack of *CYBB* nonsynonymous polymorphisms in our sample of human populations contrasts with the recurrent episodes of positive selection of the extracellular portion of gp91 during

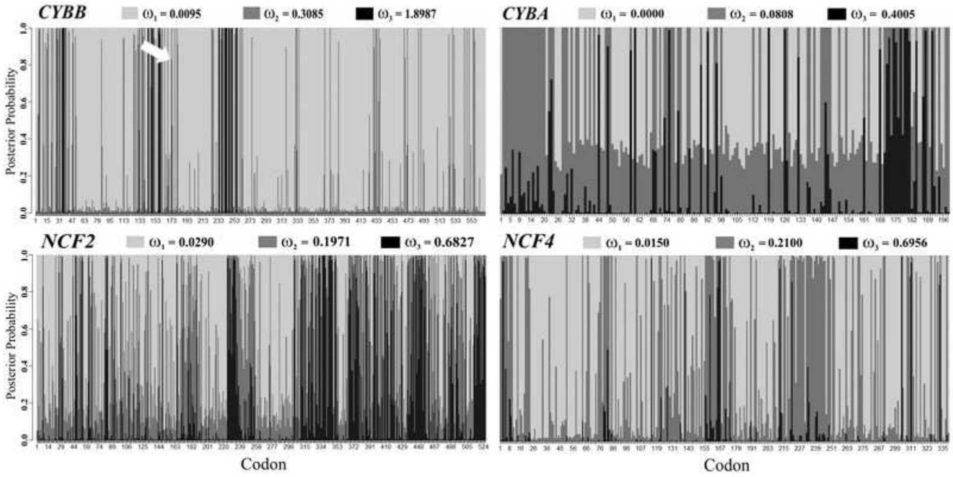
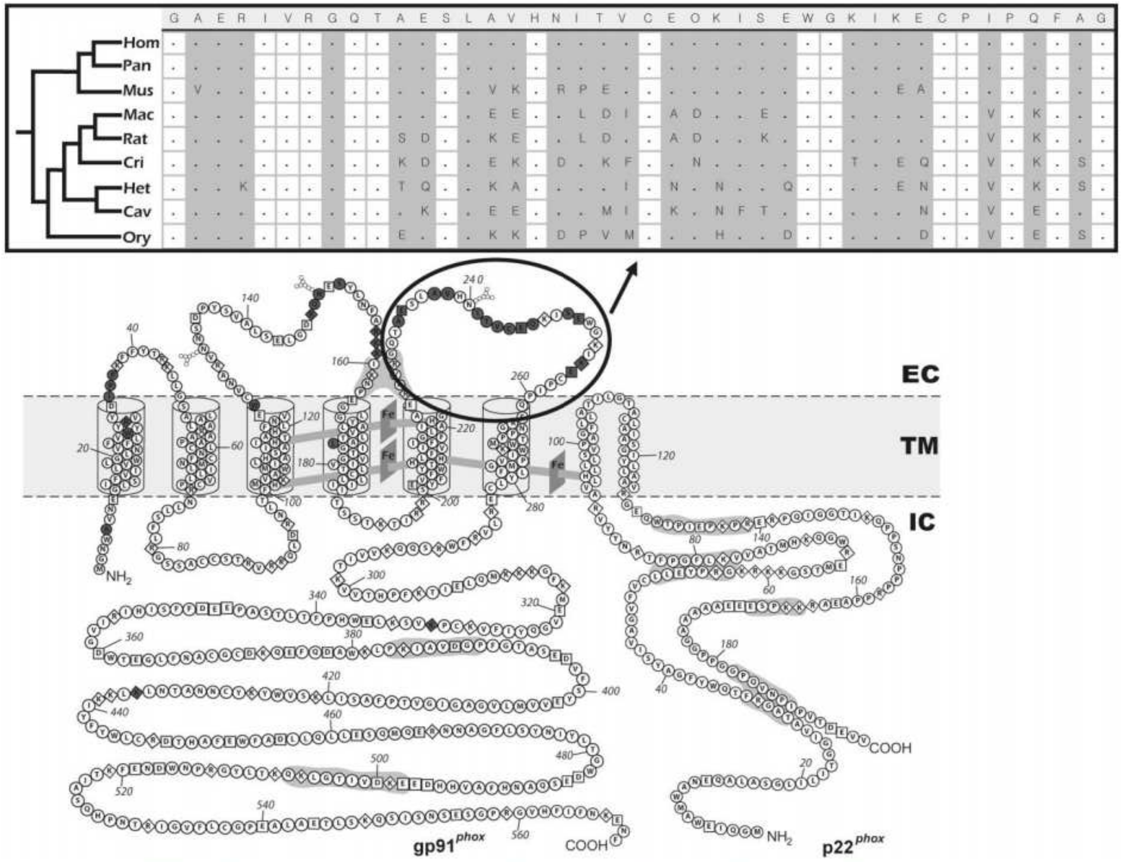


FIG. 2. Inferred types of natural selection for codons of the NADPH genes at the evolutionary time scale of mammals. Codons are represented along the horizontal axis. For each gene, three classes of sites (black, dark gray, and light gray) are considered, and each class evolved under different inferred ω values (presented for each gene in the figure at the top of each graphic). These classes correspond to the model M3 of Yang (2007a) with three classes of sites. Given our data, this model is more likely than alternative models of evolution that assume simpler scenarios, such as a unique ω for the entire gene (see supplementary material, Supplementary Material online for details regarding methods and results using alternative models). The three classes correspond to different types and levels of natural selection, from strong purifying selection (in the lightest gray) to positive selection ($\omega > 1$). For each codon, the probability of belonging to each of the three classes of ω corresponds to the height of the corresponding color in the vertical bar. For example, codon 173 of *CYBB* (indicated by a white arrow) has a 0.000 probability of belonging to the $\omega = 0.0095$ class (light gray, a class corresponding to strong purifying selection), a 0.169 probability of belonging to the $\omega = 0.3085$ class (dark gray), and a 0.831 probability of belonging to the $\omega = 1.8987$ class (black, a class that suggests positive selection). In this case, reasonable evidence of positive selection on this codon exists.

mammalian evolution, as inferred in this study. For the autosomal NADPH oxidase components (table 1 and haplotype tables online, Supplementary Material online), we observe in this study two rare and conservative nonsynonymous substitutions (i.e., involving amino acids with similar chemical properties, T85N and A304E) in *NCF4*. But for *NCF2* and *CYBA*, we observed patterns of nonsynonymous substitutions that seldom occur in human genes. *NCF2* has nine nonsynonymous substitutions; three of them are common (with a frequency higher than 5% in at least one of the studied population samples), and six are rare. On the other hand, two nonsynonymous substitutions in *CYBA* are common and ubiquitous in human populations, namely Y72H (rs4673) and V174A (rs1049254, in a position where variation among mammalian species is also observed). Moreover, the following two of the five common amino acid changes observed in the autosomal NADPH genes are predicted to be *possibly damaging* (i.e., radical) by the *Polyphen* resource (Ramensky et al. 2002); the two changes are R395W in Hispanic *NCF2* (rs13306575) and Y72H (rs4673) in *CYBA*. In general, *Polyphen* accurately predicts the effect of nonsynonymous substitutions based on biochemical and evolutionary data (Williamson et al. 2005). Notably, the Immunodeficiency Mutations Database (http://bioinf.uta.fi/NCF2base/?content=pub/IDbases, last accessed July 16, 2013) reports one 395W/395W autosomal recessive CGD patient, but we and the HapMap project (www.hapmap.org, last accessed July 16, 2013) observed the W allele at frequencies between 5% and 10% in Asians and admixed Latin Americans, including one supposedly healthy 395W/395W Japanese HapMap individual. On the basis of these results, we verified whether Native Americans, who descend from an ancestral Pleistocene Asian populations that peopled the Americas by the Behring Straits more than 14,000 years ago, may have relatively higher frequencies of this variant. We genotyped the variant 395W using a Taqman assay in 558 Native Americans (see supplementary table S3, Supplementary Material online, for detailed results) and observed an allele frequency of 1.2%, being this variant always present in heterozygous individuals.

For the four studied genes, the diversity and levels of recombination are higher in Africans than in non-Africans (table 2 and haplotype tables available as supplementary files, Supplementary Material online). This result is a consequence of the African origin of modern humans and the "out of Africa" migration that occurred 40,000–80,000 years ago after a bottleneck, leading to the peopling of other continents (Campbell and Tishkoff 2008; Laval et al. 2010). Therefore, the first divergence between continental human populations was between Africans and ancestral non-Africans. Consistently with this scenario, we observed the highest between-population differentiation for *CYBB*, *CYBA*, and *NCF4* between these two groups. Interestingly, *NCF2* does not match this pattern (table 3).



	G	A	E	R	I	V	R	G	Q	T	A	E	S	L	A	V	H	N	I	T	V	C	E	O	K	I	S	E	W	G	K	I	K	E	C	P	I	P	Q	F	A	G
Hom	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Pan	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Mus	.	V	.	.	.	.	.	.	.	.	.	.	.	.	V	K	.	R	P	E	.	.	.	.	.	.	.	.	.	.	.	.	E	A	.	.	.	.	.	.	.	.
Mac	.	.	.	.	.	.	.	.	.	.	.	.	.	.	E	E	.	.	L	D	I	.	A	D	.	.	E	.	.	.	.	.	.	.	.	.	V	.	K	.	.	.
Rat	.	.	.	.	.	.	.	.	.	.	S	D	.	.	K	E	.	.	L	D	.	.	A	D	.	.	K	.	.	.	.	.	.	.	.	.	V	.	K	.	.	.
Cri	.	.	.	.	.	.	.	.	.	.	K	D	.	.	E	K	.	D	.	K	F	.	.	N	.	.	.	.	.	.	T	.	E	Q	.	.	V	.	K	.	S	.
Het	.	.	.	K	.	.	.	.	.	.	T	Q	.	.	K	A	.	.	.	.	I	.	N	.	N	.	.	Q	.	.	.	.	E	N	.	.	V	.	K	.	S	.
Cav	.	.	.	.	.	.	.	.	.	.	.	K	.	.	E	E	.	.	.	M	I	.	K	.	N	F	T	.	.	.	.	.	.	N	.	.	V	.	E	.	.	.
Ory	.	.	.	.	.	.	.	.	.	.	E	.	.	.	K	K	.	D	P	V	M	.	.	.	H	.	.	D	.	.	.	.	.	D	.	.	V	.	E	.	S	.

Fɪɢ. 3. Natural selection mapping across *CYBB* (encoding gp91) along mammalian evolution, as identified using the PAML method by Yang (2007a). The topologies of gp91 and p22 are reproduced from Taylor et al. (2004, Copyright 2004. The American Association of Immunologists, Inc. Used with permission.). Dark gray amino acids have evolved under positive selection with >80% probability. Most of these amino acids are in the extracellular portion of the protein. The upper part of the figure shows the protein alignment for nine mammals of the gp91 region indicated by the black ellipse. In this region, a high level of amino acid variation is found between species, and several codons show $\omega > 1$. In this alignment, gray vertical bars correspond to variable amino acid sites. The protein alignment of mammals shows the following species: Hom (*Homo sapiens sapiens*), Pan (*Pan troglodytes*), Mac (*Macaca mulatta*), Mus (*Mus musculus*), Rat (*Rattus norvegicus*), Cri (*Cricetulus griseus*), Het (*Heterocephalus glaber*), Cav (*Cavia porcellus*), and Ory (*Oryctolagus cuniculus*). EC, extracelular environment; TM, transmembrane layer; and IC, intracellular environment.

Table 1. Allele Frequencies of Nonsynonymous Polymorphisms in NADPH Oxidase Genes.

Genes	rs	Minor Allele (Amino Acid)	*Polyphen* Prediction	African	European	Asian	Hispanic
CYBA							
Y72H	rs4673	T (Y)	Possibly damaging	0.46	0.32	0.17	0.22
V174A	rs1049254	T (V)	Benign	0.17	0.48	0.48	0.18
NCF2							
K181R	rs2274064	G (R)	Benign	0.35	0.37	0.41	0.48
T279M	rs13306581	T (T)	Probably damaging	0.00	0.00	0.05	0.00
V297A	rs35937854	C (A)	Benign	0.04	0.00	0.00	0.00
T361S	Chr1:181799289 NCBI36/hg18	T (S)	—	0.00	0.00	0.02	0.00
H389Q	rs17849502	A (Q)	Benign	0.00	0.05	0.00	0.07
R395W	rs13306575	T (W)	Possibly damaging	0.00	0.00	0.00	0.07
N419I	rs35012521	T (I)	Probably damaging	0.00	0.02	0.04	0.00
P454S	rs55761650	T (S)	—	0.00	0.00	0.00	0.02
L487S	Chr1:181795862 NCBI36/hg18	C (S)	—	0.02	0.00	0.00	0.00
NCF4							
T85N	rs112306225	A (N)	—	0.00	0.00	0.02	0.00
A304E	rs5995361	A (E)	Benign	0.04	0.00	0.00	0.00

From the four studied genes, *NCF4*, which encodes the regulatory protein p40-phox, shows a pattern of diversity that is typical for a gene that has evolved under neutrality. In addition to the features described in the previous paragraph, the allelic spectra of *NCF4* in the studied populations are consistent with a neutral model of evolution (tables 2–4).

Although *CYBB* presented the most interesting evolutionary history at the interspecific level, with repeated episodes of positive natural selection, recent human evolutionary history has resulted in interesting patterns of variation for *CYBA* and *NCF2*. *CYBA* encodes p22-phox, which is a transmembrane protein shared by different NADPH oxidases. In addition to harboring two common, nonsynonymous polymorphisms (V174A is also variable among different species of mammals), *CYBA* is the most variable and most affected by recombination among the NADPH oxidase genes (tables 1 and 2). In particular, *CYBA* diversity is very high in Europe: compared with 329 genes resequenced in a European sample (http://pga.gs.washington.edu/summary_stats.html, last accessed July 16, 2013), π_{CYBA} ranks 11th (i.e., the 97th percentile). Moreover, there are contrasting proportions of the total number of polymorphisms/number of singletons between Africans (a low proportion) and Europeans (a high proportion), the latter showing an excess of common polymorphisms with respect to the neutral expectation (see the D_{FL} test in table 4 and supplementary table S4, Supplementary Material online). This excess of common variants in Europeans is also significant when we conservatively tested it against a scenario of human evolution that incorporates the "Out of Africa" bottleneck (Laval et al. 2010) and the observed level of recombination in Europeans ($\rho_{CYBA} = 8.07$ for the sequenced region). Because demographic forces and recombination levels alone do not explain the high *CYBA* diversity and its excess of common polymorphisms, we suggest that balancing natural selection (that acts by maintaining different alleles at high frequency in a population) has contributed to shape the diversity of *CYBA*, at least in the European population. Indeed, figure 4*a* shows that the haplotype network of *CYBA* for the European population is consistent with the action of balancing natural selection (Bamshad and Wooding 2003), showing two well-differentiated common clades that explain the observed high diversity and the excess of common *CYBA* variants. A comparative genomic analysis confirms this inference; the ratio of polymorphisms to differences fixed between human and chimpanzee is not homogeneous along the gene in the different human populations (Mc Donald 1998; supplementary table S5, Supplementary Material online) as would be expected under neutral evolution. Our inference of balancing natural selection is consistent with the fact that 25–30% of the variation in levels of ROS production can be attributed to genetic factors (Lacy et al. 2000) and that ROS levels are associated with *CYBA* variants (Bedard et al. 2009).

p67, encoded by *NCF2*, is a necessary cytosolic NADPH component for phagocyte ROS production. Asians show a highly differentiated *NCF2* haplotype structure (see frequencies of haplotypes NCF2-D11 and NCF2-E10 in the haplotype tables online and in the network shown in fig. 4*b*), and the highest F_{ST} values are observed in pairwise comparisons between Asians and non-Asian populations (in particular with Europeans, table 3), and not between Africans and non-African populations, as is usually observed in the human genome. We confirmed these results by analyzing data for *NCF2* from the HapMap Project (supplementary material, Supplementary Material online). Moreover, a trend toward an excess of rare polymorphisms exists in Asians that is not observed elsewhere (tables 1, 2, and 4; $D_{FL} = -1.904$, $F_{FL} = -1.893$). Although we cannot exclude that this pattern of diversity is compatible with the null hypotheses of neutrality and with the tested demographic history of human populations inferred by Laval et al. (2010, tables 2–4), we can speculate and envisage four additional evolutionary scenarios

Table 2. Intrapopulation Diversity Indexes in the Studied Populations for the NADPH Oxidase Genes, Obtained from Resequencing Data[a]

	African	European	Asian	Hispanic
Number of chromosomes				
CYBB[b]	42	52	34	36
CYBA	48	62	48	46
NCF2	48	62	48	46
NCF4	48	62	48	46
Segregating sites/singletons				
CYBB	21/8	7/0	10/3	13/5
CYBA	61/22	33/3	33/5	34/7
NCF2	46/13	33/11	28/16	37/14
NCF4	45/12	26/7	19/1	30/8
Haplotype structure				
Number of inferred haplotypes[c]				
CYBB	14	5	7	12
CYBA	39	39	32	26
NCF2	38	32	18	31
NCF4	36	30	17	22
Haplotype diversity ± SD				
CYBB	0.88 ± 0.03	0.34 ± 0.08	0.53 ± 0.10	0.70 ± 0.08
CYBA	0.98 ± 0.01	0.98 ± 0.01	0.97 ± 0.00	0.96 ± 0.02
NCF2	0.99 ± 0.01	0.96 ± 0.01	0.87 ± 0.04	0.97 ± 0.01
NCF4	0.98 ± 0.01	0.93 ± 0.02	0.93 ± 0.02	0.93 ± 0.02
Recombination parameter ($\rho \times 10^3$, per site)[c]				
CYBB	0.08	<0.01	<0.01	<0.02
CYBA	2.91	1.48	0.96	0.88
NCF2	0.52	0.38	0.10	0.58
NCF4	1.37	1.03	0.39	0.22
θ estimators[d]				
$\pi \times 10^3$, per site				
CYBB	0.36	0.12	0.15	0.28
CYBA	1.90	1.63	1.51	1.57
NCF2	0.81	0.47	0.43	0.57
NCF4	1.01	0.76	0.64	0.84
$\theta_W \times 10^3$, per site				
CYBB	0.42	0.13	0.21	0.27
CYBA	2.31	1.18	1.25	1.3
NCF2	1.02	0.69	0.62	0.83
NCF4	1.01	0.70	0.54	0.87

[a] Most analyses were performed using software DnaSP (Rozas 2009).

[b] Data for *CYBB* (Xp21.1) are from Tarazona-Santos et al. (2008).

[c] Haplotypes and ρ inferred using the method by Stephens and Scheet (2005) and the software PHASE.

[d] π: Tajima (1983), θ_W: Watterson (1975).

that may have contributed to shape the pattern of *NCF2* diversity in Asians: 1) the observed trend is suggestive of a selective sweep on *NCF2* standing variation: a neutral or weakly deleterious existing variant becomes beneficial and rapidly increases in frequency (together with its associated haplotypes, i.e., incomplete sweep), reducing the nucleotide diversity in the surrounding region and rendering other standing substitutions rare. During this process, new rare substitutions appear in the expanding positively selected haplotype. 2) The pattern of diversity of *NCF2* in Asia may result from an incomplete selective sweep acting on *ARPC5*, which is located approximately 35 kb downstream of *NCF2*. In a genome-wide scan for recent positive selection, Voight et al. (2006) identified a strong signature of an incomplete sweep for *ARPC5* in Asia ($P = 0.009$), characterized by a higher than expected long-range linkage disequilibrium summarized by very high iHS statistics (see Haplotter results for the HapMap II data at http://haplotter.uchicago.edu/, last accessed July 16, 2013). SNPs in *NCF2* also presents high iHS statistics ($P = 0.02$), although values are lower than for *ARPC5*. 3) The differentiated pattern of diversity of *NCF2* in Asia may also have been generated without the action of natural selection during the first colonization of Asia by modern humans. In a process of geographic population expansion, specifically in the front wave of the expansion, some rare alleles/haplotypes (i.e., surfing alleles) may become common by chance, mimicking the pattern of diversity generated by a selective sweep (Excoffier and Ray 2008). 4) The excess of rare variants may be an artifact of pooling individuals from different populations (Ptak and Przeworski 2002). Consistent with evolutionary scenarios 1–4 that produce similar patterns of diversity, the haplotype network of *NCF2* for Eurasians (fig. 4*b*) shows the following: 1) a large differentiation between Asians and Europeans that is compatible with the high observed F_{ST} values and 2) a star-like shape associated with the haplotype NCF2-E10 that is common in Asia and rare elsewhere, which is compatible with the excess of rare alleles in the Asian populations.

Discussion

By analyzing 29 mammalian genomes and four human populations, we show in this study that natural selection has acted in different ways over time to shape the pattern of diversity of the phagocyte NADPH oxidase genes. At the temporal scale of the evolution of mammals, we have inferred recurrent episodes of positive selection acting on the extracellular portion of gp-91 that have been important to shape the pattern of interspecific diversity of this gene. Our interspecific analyses did not show a similar pattern of natural selection in any of the other phagocyte NADPH oxidase genes. Even if current knowledge on the biology of NADPH does not allow us to interpret our results in terms of function, we propose that the extracellular region of gp-91 is functionally relevant. Our results also imply that this region is highly differentiated among mammals at the protein level, and this variability should be considered when mammals models are used to study the structure and function of phagocyte NADPH components.

In the time scale of human evolution, our analyses of the NADPH oxidase genes suggest that *CYBA* has been a target of balancing natural selection. Because we do not have evidence of population-specific variants that faced selective pressure, the inferred natural selection may have acted on a standing variation in ancestral populations. This implies that the selective pressure began after the appearance of the variant and, possibly, acted in a specific geographic region (Barret and Schluter 2008). The signatures of natural selection acting on a new mutation and on standing variation differ. In the case of selective sweeps, episodes of natural selection on standing variation are associated to a larger variance in the allelic spectrum with respect to natural selection on a new mutation. Also, selection on standing variation may produce an excess of alleles at intermediate frequencies that is not associated with high nucleotide diversity (Przeworski et al. 2005; Peter et al. 2012). This pattern contrasts with the effect of balancing

Table 3. Pairwise F_{ST} Genetic Distances between Populations.

	CYBB[a]				CYBA			
	Africa	Europe	Asia	Hispanic	Africa	Europe	Asia	Hispanic
Africa	—	0.316	0.264	0.092	—	0.074	0.083	0.065
Europe	0.257	—	0.000	0.107		—	0.002	0.056
Asia	0.211	0.000	—	0.073			—	0.054
Hispanic	0.070	0.082	0.056	—				—
	NCF2				NCF4			
Africa	—	0.048	0.058	0.037	—	0.128	0.136	0.158
Europe		—	0.069	0.000		—	0.026	0.026
Asia			—	0.059			—	0.005
Hispanic				—				—

[a] For *CYBB* F_{ST} estimators are above the diagonal. Below the diagonal are the F_{ST} values corrected as if the effective population sizes of X chromosome genes were equal to autosomal ones.

Table 4. Results of Neutrality Tests for the NADPH Oxidase Genes and Their Significance.[a]

	African	European	Asian	Hispanic
Tajima's *D*				
CYBB	−0.473	−0.274	−0.813	0.084
CYBA	−0.580	1.242	0.684	0.707
NCF2	−0.412	−0.939	−0.883	−0.987
NCF4	−0.750	0.243	0.556	−0.102
Fu and Li's *D*				
CYBB	−1.050	1.110	0.395	−0.977
CYBA	−1.485	1.734*	1.188	0.138
NCF2	−0.407	−1.105	−1.904	−1.398
NCF4	−0.308	−0.443	−1.893	−0.266
Fu and Li's *F*				
CYBB	−0.980	0.811	0.114	−0.814
CYBA	−1.382	1.893*	1.205	0.405
NCF2	−0.512	−1.252	−1.893	1.567
NCF4	−0.557	−0.231	1.225	−0.248

NOTE.—Underlined values represent significant results under the demographic model inferred by Laval et al. (2010) for human populations. See details in supplementary table S4, Supplementary Material online.

[a] The McDonald–Kreitman test is nonsignificant in any of the cases.

*Significant under the Wright–Fisher model of constant population size.

natural selection, which produces an excess of common alleles associated with high genetic diversity. Thus, the observed pattern of *CYBA* diversity in Europeans is not consistent with a selective sweep on a standing variation, but it is consistent with a scenario of balancing selection acting on standing variation.

If we consider for *CYBA* that heterozygote advantage may be the mechanisms of balancing selection, we can speculate that the biological basis for this mechanism may be the following: considering that p22-phox is not exclusive of the phagocyte NADPH oxidase, but it is also part of Nox complexes expressed in other tissues, the dependence of ROS production on *CYBA* variants has to be finely regulated. If *CYBA* variants induce high levels of ROS, these variants may favor a phagocyte-dependent efficient response to pathogens but may damage other endothelial tissues. Alternatively, tissue oxidative damage does not occur if ROS production is low, but this response may be associated with a weaker phagocyte respiratory burst against pathogens. In this context, heterozygote individuals with a *CYBA*-dependent intermediate level of ROS production may have been favored by natural selection.

Our results contribute to the discussion regarding the relevance of balancing selection in shaping the diversity of innate immunity genes (Ferrer-Admetlla et al. 2008; Barreiro et al. 2009). Ferrer-Admetlla et al. (2008) have associated the recurrent signatures of balancing selection on inflammatory genes with the need for fine regulation. *CYBA* is the only phagocytic NADPH oxidase gene that also encodes nonphagocyte Nox components; thus, *CYBA* has a role in ROS cell signaling, a potentially dangerous process due to its capability to produce oxidative damage to tissues. Genes with these characteristics likely need even tighter regulation. The interplay between pathogen-driven selective pressure on innate immunity genes and their concomitant nonimmunological functions is complex. In addition to *CYBA*, other interesting examples of this interplay can be found among the 10 human toll-like receptors (TLRs) that show a variety of signatures of natural selection, *TLR8*, which shows the strongest signature of purifying selection, is also involved in neuronal development (Barreiro et al. 2009), and it is difficult to discriminate the role of each function in determining the observed signature of natural selection.

With few exceptions, the pathogens responsible for natural selection on immune genes are difficult to specify. In the case of NADPH oxidase, we can infer, based on the spectrum of infections in CGD patients, that catalase-positive bacteria and fungi, such as *Staphylococci*, *Salmonella*, *Candida*, *Aspergillus*, and *M. tuberculosis*, may be the selective pathogens. Interactions between the host and pathogens also include mechanisms of the latter to impair the respiratory burst of the former. For example, *Leishmania donovani* blocks the assembly of NADPH oxidase at the phagosome membrane (Lodge et al. 2006). These mechanisms may constitute selective pressures imposed by pathogens.

The associations reported in GWAS between rs4821544 in *NCF4* and Crohn's disease (an idiopathic inflammatory bowel disease that predominantly involves the ileum and colon, Rioux et al. 2007) and between rs10911363 in *NCF2* and

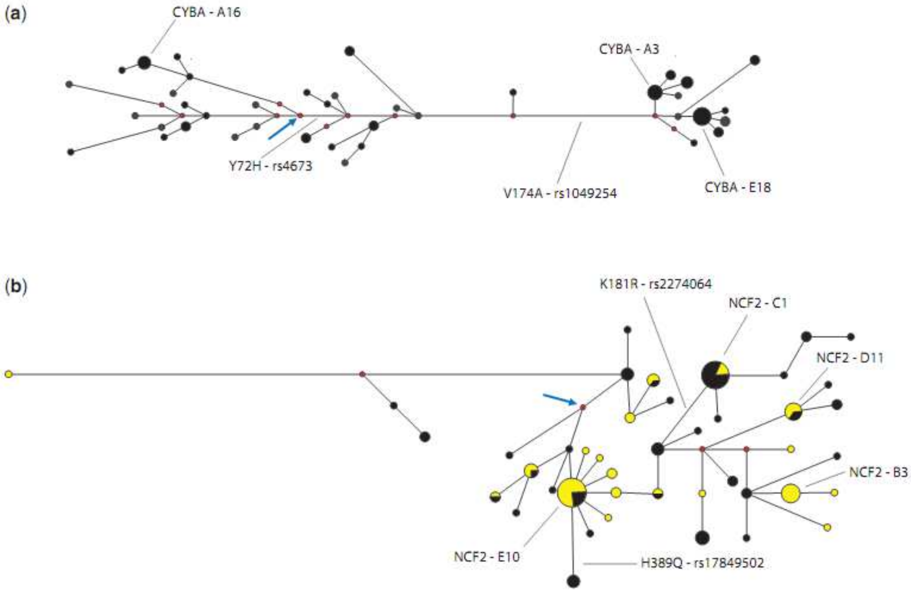


FIG. 4. Phylogenetic networks of (*a*) *CYBA* in Europeans and (*b*) *NCF2* in Europeans (black) and Asians (yellow). The lengths of the branches are proportional to the number of mutations. Only nonsynonymous mutations are shown. The haplotype names correspond to the table of inferred haplotypes in the supplementary material, Supplementary Material online. We only show the names of haplotypes with a frequency >5%. Ancestral haplotypes (inferred as being the human haplotype or median vector most similar to the chimpanzee sequence) are indicated by an arrow. Median vectors are in red.

systemic lupus erythematosus (Cunninghame Graham et al. 2011) confirm the involvement of NADPH genes in the pathogenesis of inflammatory-related common diseases. Our claim that natural selection acted on *CYBA* (and maybe in *NCF2*) is relevant for biomedical studies because combining evidence of natural selection with association analyses in immune genes increases the statistical power to detect disease-associated variants (Ayodo et al. 2007). As a new generation of association studies focusing on rare variation is emerging, the combination of genes deemed interesting from GWAS and populations with an excess of rare variants in these genes, such as *NCF2* in Asians, are particularly interesting as a source of rare variants with clinical relevance. Finally, by determining through molecular evolution mapping that the extracellular portion of gp91 (encoded by *CYBB* in the X-chromosome) has been subject to recurrent episodes of positive selection at the scale of mammals evolution, we posit the hypothesis that this portion of the NADPH oxidase is relevant for currently unknown biological processes that, once revealed by structural and functional investigations, will contribute to understanding the role of NADPH oxidase in infectious, autoimmune, and cardiovascular diseases.

Materials and Methods

Molecular Evolution Analysis of NADPH Genes

We used the maximum likelihood framework developed by Yang (2007a) to estimate ω for the NADPH oxidase genes under a variety of evolutionary models, as implemented in the PAML software (Yang 2007b). This approach allows inferences about the evolution of a coding region along an inter-specific phylogeny, mapping the codons that have evolved under strong/weak purifying selection, neutrality, or adaptive positive selection. Further details about these analyses are available as supplementary material, Supplementary Material online.

Human Population Genetics of NADPH Genes

For human population genetics analyses, we conducted bidirectional Sanger sequencing of *CYBB*, *CYBA*, *NCF2*, and *NCF4* for a total of 35,242 bp for each of 102 healthy individuals as part of the SNP500 Cancer project (Packer et al. 2006; see supplementary fig. S1, Supplementary Material online, for details). Human population resequencing data for *CYBB* were published in Tarazona-Santos et al. (2008). The resequencing experiments were performed as in Packer et al. (2006). These 102 unrelated individuals include the following: 24 of African ancestry (15 African Americans from the United States and 9 Pygmies), 23 admixed Latin Americans (from Mexico, Puerto Rico, and South America), 31 Europeans (from the CEPH/UTAH pedigree and the NIEHS Environmental Genome Project), and 24 Asians–Oceanians (from Melanesia, Pakistan, China, Cambodia, Japan, and Taiwan).

After controlling for multiple tests, we confirmed that all SNPs fit the Hardy–Weinberg equilibrium by the Guo and

Thompson (1992) test, which was implemented in the software Arlequin 3.0 (Excoffier et al. 2007). Insertion-deletions (INDELs) were excluded from population genetics analyses. To assess intrapopulation diversity, we used two statistics: π, the per-site mean number of pairwise differences between sequences (Tajima 1983), and θ_w, based on the number of segregating sites (S) (Watterson 1975). We measured pairwise between-populations diversity by using the F_{ST} statistics calculated using the software DnaSP (Rozas 2009).

Haplotypes and the recombination parameter ρ were inferred using the PHASE software (Stephens and Scheet 2005), and diversity indexes calculations (tables 2 and 3) as well as neutrality statistics (table 4) were estimated using DnaSP software. We applied two kinds of neutrality tests: 1) tests based on the allelic spectrum, which is the distribution of polymorphisms across different classes of frequencies, namely, Tajima's D_T (Tajima 1989) and Fu–Li's D_{FL} and F_{FL} (Fu and Li 1993) and 2) tests based on comparisons between the number of polymorphisms in human populations and fixed differences with the chimpanzee (i.e., outgroup), namely, the McDonald and Kreitman (1991) test and the adapted Kolmogorov–Smirnoff test by McDonald (1998). For the first set of tests, we used as null hypotheses both the classic Wright–Fisher model of neutrality with a constant population size, as well as the more realistic evolutionary scenario for human populations inferred by Laval et al. (2010). In the case of the scenario of Laval et al. (2010), we ignored intercontinental gene flow within the Old World because these rare gene flow events likely does not affect the level of significance of the neutrality tests given its very low inferred values (1.3×10^{-5}). Null distributions used to test the significance of the neutrality tests under these evolutionary scenarios were generated using coalescent simulations and a significance level of 0.05 (Hudson 2002). The Kolmogorov–Smirnoff test of neutrality adapted by McDonald (1998) was performed using Slider software available at http://udel.edu/~mcdonald/aboutdnaslider.html (last accessed July 16, 2013). We performed coalescent simulations using ms software (Hudson 2002). Further methodological details are available as supplementary material, Supplementary Material online. We constructed the *CYBA* and *NCF2* networks using all SNP variants and applying the Median joining algorithm and the maximum parsimony option calculations as implemented in the software Network 4.6 (Bandelt et al. 1999).

Supplementary Material

Supplementary tables S1–S5 and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

S.J.C. conceived the study; E.T.-S., M.M., F.L., R.C., A.C., C.F., L.P., and B.S. generated the data/analyzed the sequences; L.B., A.C., W.C.S.M., and M.Y. provided tools for data generation and analyses; E.T.-S., M.M., F.L., W.C.S.M., and R.R. performed population genetics analyses; E.T.-S. and M.M. prepared tables and figures; E.T.-S. and S.J.C. wrote the manuscript. All the authors contributed with discussions about the results. The authors are grateful to Silvia Fuselli for discussions of our results, to Fernando Levi Soares for technical assistance with the figures and the Sequencing Group of the Core Genotyping Facility (National Cancer Institute) for their technical assistance. This work was supported by the Intramural Research Program of the National Cancer Institute (NCI) and National Institutes of Health (NIH). C.F. was supported by the University of Bologna. E.T-S., M.M., W.C.S.M., and F.L. were supported by the Fogarty International Center and National Cancer Institute (1R01TW007894), Conselho Nacional de Desenvolvimento Científico e Tecnológico (Brazil, Grant 307556/2012-3), Fundação de Amparo a Pesquisa de Minas Gerais (Brazil, CBB PPM 00265/12) and the Brazilian Ministry of Education (CAPES Agency, grant PNPD 02567/09-1).

References

Ayodo G, Price AL, Keinan A, Ajwang A, Otieno MF, Orago AS, Patterson N, Reich D. 2007. Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am J Hum Genet.* 81(2):234–242.

Bamshad M, Wooding SP. 2003. Signatures of natural selection in the human genome. *Nat Rev Genet.* 4(2):99–111.

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16(1):37–48.

Barreiro LB, Ben-Ali M, Quach H, et al. (17 co-authors). 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5(7):e1000562.

Barreiro LB, Quintana-Murci L. 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 11(1):17–30.

Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol Evol.* 23(1):38–44.

Bedard K, Attar H, Bonnefont J, Jaquet V, Borel C, Plastre O, Stasia MJ, Antonarakis SE, Krause KH. 2009. Three common polymorphisms in the CYBA gene form a haplotype associated with decreased ROS generation. *Hum Mutat.* 30(7):1123–1133.

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M. 2008. Natural selection on genes that underlie human disease susceptibility. *Curr Biol.* 18(12):883–889.

Brandes RP, Kreuzer J. 2005. Vascular NADPH oxidases: molecular mechanisms of activation. *Cardiovasc Res.* 65(1):16–27.

Buckley RH. 2004. Pulmonary complications of primary immunodeficiencies. *Paediatr Respir Rev.* 5(Suppl A):S225–S233.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (13 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437(7062):1153–1157.

Bustamante J, Arias AA, Vogt G, et al. (29 co-authors). 2011. Germline CYBB mutations that selectively affect macrophages in kindreds with X-linked predisposition to tuberculous mycobacterial disease. *Nat Immunol.* 12(3):213–221.

Campbell MC, Tishkoff SA. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet.* 9:403–433.

Chanock SJ, Roesler J, Zhan S, Hopkins P, Lee P, Barrett DT, Christensen BL, Curnutte JT, Görlach A. 2000. Genomic structure of the human p47-phox (NCF1) gene. *Blood Cells Mol Dis.* 26(1):37–46.

Cunninghame Graham DS, Morris DL, Bhangale TR, Criswell LA, Syvänen AC, Rönnblom L, Behrens TW, Graham RR, Vyse TJ. 2011. Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with systemic lupus erythematosus. *PLoS Genet.* 7(10):e1002341.

Excoffier L, Laval G, Schneider S. 2007. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol Bioinform Online.* 1:47–50.

Excoffier L, Ray N. 2008. Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol Evol.* 23(7):347–351.

Ferrer-Admetlla A, Bosch E, Sikora M, Marquès-Bonet T, Ramírez-Soriano A, Muntasell A, Navarro A, Lazarus R, Calafell F, Bertranpetit J, Casals F. 2008. Balancing selection is the main force shaping the evolution of innate immunity genes. *J Immunol.* 181(2):1315–1322.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133(3):693–709.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.

The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

Guo SW, Thompson EA. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48(2):361–372.

Heyworth PG, Cross AR, Curnutte JT. 2003. Chronic granulomatous disease. *Curr Opin Immunol.* 15(5):578–584.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Jacob CO, Eisenstein M, Dinauer MC, et al. (21 co-authors). 2012. Lupus-associated causal mutation in neutrophil cytosolic factor 2 (NCF2) brings unique insights to the structure and function of NADPH oxidase. *Proc Natl Acad Sci U S A.* 109(2):E59–E67.

Kimura M. 1974. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4(8):e1000144.

Lacy F, Kailasam MT, O'Connor DT, Schmid-Schönbein GW, Parmer RJ. 2000. Plasma hydrogen peroxide production in human essential hypertension: role of heredity, gender, and ethnicity. *Hypertension* 36(5):878–884.

Laval G, Patin E, Barreiro LB, Quintana-Murci L. 2010. Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS One* 5(4):e10284.

Lewontin R. 1972. The apportionment of human diversity. *Evol Biol.* 6: 391–398.

Lindblad-Toh K, Garber M, Zuk O, et al. (88 co-authors). 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–482.

Lodge R, Diallo TO, Descoteaux A. 2006. *Leishmania donovani* lipophosphoglycan blocks NADPH oxidase assembly at the phagosome membrane. *Cell Microbiol.* 8(12):1922–1931.

Magalhães WC, Rodrigues MR, Silva D, Soares-Souza G, Iannini ML, Cerqueira GC, Faria-Campos AC, Tarazona-Santos E. 2012. DIVERGENOME: a bioinformatics platform to assist population genetics and genetic epidemiology studies. *Genet Epidemiol.* 36(4):360–367.

Marth JD, Grewal PK. 2008. Mammalian glycosylation in immunity. *Nat Rev Immunol.* 8(11):874–887.

McDonald JH. 1998. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol.* 15:377–384.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila. Nature* 351(6328):652–654.

Nielsen R, Bustamante C, Clark AG, et al. (12 co-authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* 3(6):e170.

Olsson LM, Lindqvist AK, Källberg H, Padyukov L, Burkhardt H, Alfredsson L, Klareskog L, Holmdahl R. 2007. A case-control study of rheumatoid arthritis identifies an associated single nucleotide polymorphism in the NCF4 gene, supporting a role for the NADPH-oxidase complex in autoimmunity. *Arthritis Res Ther.* 9(5):R98.

Packer BR, Yeager M, Burdett L, et al. (13 co-authors). 2006. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.* 34(Database issue):D617–D621.

Peter BM, Huerta-Sanchez E, Nielsen R. 2012. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS Genet.* 8(10):e1003011.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312–2323.

Ptak SE, Przeworski M. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* 18(11):559–563.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30(17):3894–3900.

Rioux JD, Xavier RJ, Taylor KD, et al. (24 co-authors). 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet.* 39(5):596–604.

Roberts RL, Hollis-Moffatt JE, Gearry RB, Kennedy MA, Barclay ML, Merriman TR. 2008. Confirmation of association of IRGM and NCF4 with ileal Crohn's disease in a population-based cohort. *Genes Immun.* 9(6):561–565.

Rozas J. 2009. DNA sequence polymorphism analysis using DnaSP. *Methods Mol Biol.* 537:337–350.

San José G, Fortuño A, Beloqui O, Díez J, Zalba G. 2008. NADPH oxidase CYBA polymorphisms, oxidative stress and cardiovascular diseases. *Clin Sci (Lond).* 114(3):173–182.

Santiago HC, Gonzalez Lombana CZ, Macedo JP, et al. (12 co-authors). 2012. NADPH phagocyte oxidase knockout mice control *Trypanosoma cruzi* proliferation, but develop circulatory collapse and succumb to infection. *PLoS Negl Trop Dis.* 6(2):e1492.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet.* 76(3):449–462.

Sumimoto H, Miyano K, Takeya R. 2005. Molecular composition and regulation of the Nox family NAD(P)H oxidases. *Biochem Biophys Res Commun.* 338(1):677–686.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105(2):437–460.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.

Tarazona-Santos E, Bernig T, Burdett L, Magalhaes WC, Fabbri C, Liao J, Redondo RA, Welch R, Yeager M, Chanock SJ. 2008. CYBB, an NADPH-oxidase gene: restricted diversity in humans and evidence for differential long-term purifying selection on transmembrane and cytosolic domains. *Hum Mutat.* 29(5):623–632.

Taylor RM, Burritt JB, Baniulis D, Foubert TR, Lord CI, Dinauer MC, Parkos CA, Jesaitis AJ. 2004. Site-specific inhibitors of NADPH oxidase activity and structural probes of flavocytochrome b: characterization of six monoclonal antibodies to the p22phox subunit. *J Immunol.* 173(12):7349–7357.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4(3):e72.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7(2): 256–276.

Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A.* 102(22):7882–7887.

Yang Z. 2007a. Adaptive molecular evolution. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of statistical genetics, Vol. 1, 3rd ed. Susex (United Kingdom): John Wiley & Sons. p. 377–406.

Yang Z. 2007b. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.