Universidade de São Paulo
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Departamento de Química
Programa de Pós-Graduação em Química

“Análise de marcadores forenses (STRs e SNPs) rotineiramente empregados na identificação humana utilizando sequenciamento de nova geração”.

Guilherme do Valle Silva

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: **Química**

RIBEIRÃO PRETO -SP

2018

GUILHERME DO VALLE SILVA

**Análise de marcadores forenses (STRs e SNPs) rotineiramente empregados na
identificação humana utilizando sequenciamento de nova geração**

Versão corrigida

Dissertação apresentada à Faculdade de
Filosofia, Ciências e Letras de Ribeirão Preto da
Universidade de São Paulo, como parte das exigências
para a obtenção do título de Mestre em Ciências.

Área de concentração: Química

Orientador: Prof. Dr. Celso Teixeira Mendes Junior

RIBEIRÃO PRETO - SP

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Valle-Silva, Guilherme

Análise de marcadores forenses (STRs e SNPs) rotineiramente empregados na identificação humana utilizando sequenciamento de nova geração. Ribeirão Preto, 2018.

151 p. : il. ; 30cm

Dissertação de Mestrado, apresentada à Faculdade de Filosofia Ciências e Letras de Ribeirão Preto/USP – Área de concentração: Química.

Orientador: Mendes-Junior, Celso Teixeira.

1. Genética Forense. 2. Sequenciamento de Nova Geração. 3.SNPs. 4. STRs. 5. Ancestralidade.

Nome: Silva, Guilherme do Valle

Título: Análise de marcadores forenses (STRs e SNPs) rotineiramente empregados na identificação humana utilizando sequenciamento de nova geração

Dissertação apresentada à Faculdade de
Filosofia, Ciências e Letras de Ribeirão Preto da
Universidade de São Paulo, como parte das exigências
para a obtenção do título de Mestre em Ciências, Área:
Química

Aprovado em: 05/10/2018

Banca Examinadora:

Prof. Dr. Rodrigo Soares Moura Neto
Instituição: Universidade Federal do Rio de Janeiro
Julgamento: Aprovado
Assinatura:

Prof. Dr. Aline Thaís Bruni
Instituição: Faculdade de Filosofia Ciências e Letras de Ribeirão Preto
Julgamento: Aprovado
Assinatura:

Prof. Dr. Caio César de Cerqueira
Instituição: Polícia Técnico-Científica do Estado de São Paulo
Julgamento: Aprovado
Assinatura:

Orientador: Prof. Dr. Celso Teixeira Mendes Junior
Assinatura:

Aos meus pais e meu irmão, por todo apoio e carinho

Agradecimentos

Agradeço primeiramente a Deus por toda força e pelas oportunidades dadas a mim, não só para realização desse trabalho, mas também por tudo que tenho na vida.

À toda minha família que sempre apoia e incentiva os meus sonhos. Em especial meus pais: Dimas e Nelma, e meu irmão Antônio, que me dão todo amparo, cuidado e amor necessários para que eu possa sempre seguir em frente com meus objetivos.

Ao meu orientador, Professor Celso, que me inspira, tanto pelo profissional, quanto pela pessoa que é. Sem dúvidas foi o peso que faltava para que eu seguisse no mestrado, ao me apresentar uma área tão linda como a Genética Forense.

À minha namorada Andressa, pela paciência e carinho que teve por mim durante esse período, aguentando minhas queixas e preocupações. Sei que não foi fácil!!!

Agradeço aos meus amigos de Votuporanga (alguns aguento desde a pré-escola) e aos meus amigos que Ribeirão e a USP me trouxeram. Mesmo que às vezes distante, a amizade sempre continua igual.

Ao nosso grupo do Laboratório de Pesquisas Forenses e Genômicas, Letícia, Alison, Malu, Nádia, Gui D, Thássia, Amanda e André por todos os momentos vividos dentro e fora do laboratório. Momentos esses que foram de diversões, partilha de preocupações, conversas ativas e ajudas (muita ajuda!!).

À Universidade de São Paulo, pela infraestrutura. Bem como aos professores de Graduação e Pós-Graduação, funcionários e à Coordenação do Programa de Pós-Graduação em Química, pelo aprendizado e apoio.

Ao Prof. Dr. Aguinaldo Luiz Simões, por ceder à utilização dos equipamentos.

Ao Laboratório de Imunopatologia Molecular, na figura do Prof. Dr. Eduardo Antônio Donadi, pela infraestrutura para realização deste trabalho.

Por fim, agradeço a cada voluntário pela disposição e paciência, pois sem eles não haveria pesquisa.

Resumo

VALLE-SILVA, G. **Análise de marcadores forenses (STRs e SNPs) rotineiramente empregados na identificação humana utilizando sequenciamento de nova geração.** 2018. 154 páginas. Dissertação de Mestrado – Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, 2018.

A genética forense vem se desenvolvendo cada vez mais, com novas tecnologias e implementação de novos conjuntos de marcadores de DNA com maiores níveis de informatividade. Os marcadores genéticos são amplamente usados na identificação humana, pois permitem distinguir indivíduos com alta acurácia. Duas classes de marcadores muito utilizadas atualmente são os STRs (*Short Tandem Repeats*) e os SNPs (*Single Nucleotide Polymorphisms*). Os STRs são altamente informativos e, portanto, úteis para a prática forense. Kits mais novos como *GlobalFiler* (*Thermo Fisher Scientific*) e *PowerPlex Fusion System* (*Promega*) apresentam a análise de mais de 20 *loci* STRs de uma só vez. Já os SNPs, por possuírem sua informatividade mais reduzida (necessita de mais *loci* analisados), são menos utilizados, porém apresentam vantagem em amostras degradadas de DNA; assim, conjuntos de identificação como o *52-plex* desenvolvido pelo consórcio *SNPforID* e o conjunto *IISNPs*, vêm sendo estudados em várias populações do mundo. Com o desenvolvimento de técnicas de sequenciamento de nova geração (NGS – *Next Generation Sequencing*) para análise de DNA, a obtenção de perfis de DNA se tornou mais acurada. Algumas plataformas permitem gerar perfis de até 96 indivíduos simultaneamente. Este estudo tem por objetivo principal analisar 171 marcadores genéticos (Amelogenina, Y-INDEL, 30 STRs e 139 SNPs) em 340 indivíduos miscigenados da região da cidade de Ribeirão Preto (SP) utilizando a plataforma de sequenciamento de nova geração *MiSeq Personal Sequencer* (*Illumina Inc.*), bem como calcular as frequências alélicas e genotípicas, verificar a aderência ao equilíbrio de Hardy-Weinberg e estimar parâmetros forenses para os diferentes conjuntos de marcadores. Análises de ancestralidade foram realizadas para os conjuntos de SNPs. Para o preparo das bibliotecas de amostras a serem sequenciadas, foi utilizado o kit *HaloPlex* (*Agilent Technologies, Inc*), onde foram incluídos os marcadores dos kits *GlobalFiler* e *PowerPlex Fusion System*, e os SNPs existentes no conjunto do consórcio *SNPforID* (*52-plex*) e *IISNPs* (92 SNPs). De todos os marcadores incluídos no ensaio, apenas um SNP (rs763869) presente no conjunto *SNPforID* não pôde ser analisado devido a questões técnicas. Dos 139 SNPs analisados apenas seis apresentaram desvios significativos em relação ao equilíbrio de Hardy-Weinberg,

número este esperado devido ao acaso. Os conjuntos de SNPs apresentam elevada informatividade com Probabilidade de *Match* de $6,48 \times 10^{-21}$ (*52-plex*) a $4,91 \times 10^{-38}$ (*IISNP*), e Poder de Exclusão de 0,9997 (*52-plex*) e 0,99999997 (*IISNP*). De modo geral, as inferências de ancestralidade obtida utilizando estes conjuntos, indicaram elevada contribuição europeia (superior a 70%) e baixa contribuição ameríndia (inferior a 10%) na população, enquanto que as análises de mistura individual se mostraram consistentes, com a maioria dos indivíduos apresentando elevada ancestralidade europeia. Os resultados dos marcadores relativos ao sexo (Amelogenina, Y-INDEL e DYS391) foram consistentes com o sexo dos doadores das amostras. As frequências alélicas e parâmetros forenses foram calculados para os STRs, revelando uma alta informatividade. A Probabilidade de *Match* combinada e o Poder de Exclusão combinado foram de $1,19 \times 10^{-36}$ e 0,999999999997 respectivamente. Dos 29 STRs autossômicos presentes, seis apresentaram desvios ao equilíbrio de Hardy-Weinberg, refletindo possíveis falhas no sequenciamento e genotipagem destes marcadores.

Palavras-Chave: Genética forense; sequenciamento de nova geração; Single Nucleotide Polymorphisms; Short Tandem Repeats; ancestralidade.

Abstract

VALLE-SILVA, G. **Analysis of forensic markers (STRs and SNPs) routinely used in human identification assays by means of next generation sequencing**. 2018. 154 pages. Dissertation (Master's Degree) - Faculty of Philosophy, Sciences and Letters at Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, 2018.

The field of forensic genetics has developed increasingly with the implementation of new sets of DNA markers with higher levels of informativeness. The genetic markers are widely used in human identification as they allow distinguishing individuals with high accuracy. Two of the most commonly used markers are the Short Tandem Repeats (STRs) and the Single Nucleotide Polymorphisms (SNPs). Newer kits such as *GlobalFiler* (Thermo Fisher Scientific) and *PowerPlex Fusion System* (Promega) can analyze more than 20 STRs *loci* at once. When comparing with STRs, the SNPs are less informative and many more *loci* are needed to reach the same informativeness of STR kits. However, they are advantageous when using degraded DNA samples. The identification sets such as the 52-plex developed by the SNPforID Consortium and the *IISNPs* have been analyzed in many worldwide populations. With the development of next generation sequencing techniques (NGS– *Next Generation Sequencing*), obtaining DNA profiles has become more accurate and some platforms allow generating profiles of up to 96 individuals simultaneously. The main goal of this study is to analyze 171 markers (Amelogenin, Y-INDEL, 30 STRs and 139 SNPs) in 340 admixed individuals from Ribeirão Preto, SP, using the NGS platform MiSeq Personal Sequencer (Illumina Inc.). This will allow the calculation of allele and genotype frequencies, the verification of adherence to Hardy-Weinberg's equilibrium and the estimation of forensic parameters for each set of marker. Ancestry analysis was performed for the sets of SNPs. The *HaloPlex* kit (Agilent Technologies, Inc) was used for library preparation including the STRs from the kits *GlobalFiler* and *PowerPlex Fusion System* and the SNPs from the SNPforID consortium (52-plex) and *IISNPs* (92 SNPs) identification sets. A single SNP (rs763869) from the SNPforID set was not analyzed due to technical issues. Only six of the 139 analyzed SNPs presented significant deviation from the Hardy-Weinberg equilibrium expectations, which is expected by chance alone. The SNPs sets exhibited high informativeness, with match

probability ranging from 6.48×10^{-21} (52-plex) to 4.91×10^{-38} (*IISNPs*) and exclusion power of 0.9997 (52-plex) and 0.99999997 (*IISNPs*). In general, ancestry estimates obtained using these sets indicated a high European contribution (higher than 70%) and low Amerindian contribution (less than 10%) in the population sample, while the individual admixture analyses exhibited were highly consistent, with the majority of individuals presenting high European ancestry. The results of the sex markers (Amelogenin, Y-INDEL and DYS391) were in agreement with the reported sexes from sample donors. The allele frequencies and forensic parameters calculated for the STRs revealed high informativeness. The combined match probability and the combined exclusion power were 1.19×10^{-36} and 0.999999999997 respectively. Six of the 29 autosomal STRs presented significant deviations from the Hardy-Weinberg equilibrium expectations, reflecting possible failures in sequencing and genotyping of these markers.

Key words: Forensic genetics; Next generation sequencing; Single Nucleotide Polymorphisms; Short Tandem Repeats; Ancestry.

Lista de Figuras

- Figura 1.** Relação de kits para obtenção de perfis de STRs presentes no mercado. Em amarelo destaca-se pares de *loci* que estão no mesmo cromossomo. Os *loci* presentes em cada kit estão indicados pelo preenchimento da célula nas colunas pertinentes de cores cinzas. As células preenchidas pela cor preta, indicam os *loci* que devem ser analisados pelo sistema CODIS, ESS (conjuntos de STRs analisados na Europa) e Interpol (adaptado de Butler et al. 2012). . 21
- Figura 2.** Representação geral do preparo da biblioteca utilizando-se o kit HaloPlex (Adaptado de <https://www.agilent.com/cs/library/usermanuals/Public/G9900-90001.pdf>). ... 33
- Figura 3.** Diagrama triangular gerado pelo *software* STRUCTURE para os conjuntos SNPforID 52-plex e IISNPs, representando as estimativas de ancestralidade geradas para cada indivíduo. Cada indivíduo é representado por um círculo colorido, sendo em amarelo os indivíduos da população de Ribeirão Preto -SP. Indivíduos das populações parentais estão mais concentrados nos vértices do triângulo: EUR – Europeus (vermelhos); AMR – Ameríndios (SNPforID)/Asiáticos (1000Genomes) (azuis); AFR – Africanos (verdes). 51
- Figura 4.** Gráfico retangular gerado pelo *software* STRUCTURE representando as estimativas de ancestralidade geradas para cada indivíduo. Cada indivíduo é representado por uma barra que pode apresentar até três cores, sendo que vermelho, verde e azul correspondem a ancestralidade europeia (EUR), africana (AFR) e ameríndia/asiática (AMR/EAS), respectivamente. 52
- Figura 5.** Gráfico da porcentagem de sucesso de genotipagem dos 31 *loci* por cada *software* individualmente e após junção dos dados gerando genótipos finais. 55
- Figura 6.** Gráfico da porcentagem de genótipos finais obtidos com sucesso e quantidade de amostras que alcançaram tal porcentagem de sucesso. 56
- Figura 7.** Cobertura média de cada *locus* observada nos dois *softwares* empregados. As barras pretas representam o desvio padrão encontrado. 57
- Figura 8.** Gráfico da razão de *stutter* para cada *locus* de STR. 62
- Figura 9.** Imagem do *software* IGV indicando o alinhamento das *reads* na região do gene da Amelogenina do cromossomo X. Na parte superior (A) está representada amostra do sexo masculino; a parte inferior (B) representa amostra do sexo feminino..... 65
- Figura 10.** Imagem do *software* IGV indicando o alinhamento das *reads* na região do gene da Amelogenina do cromossomo Y. Na parte superior (A) está representada amostra do sexo masculino; a parte inferior (B) representa amostra do sexo feminino..... 65
- Figura 11.** Imagem do *software* IGV indicando a região do *locus* TPOx na amostra PIG0002, a qual possui genótipo 9/11. O genoma de referência (retângulo amarelo) apresenta alelo composto por 8 unidades de repetição. O quadro preto delimita a região repetitiva, e suas regiões flangeadoras à sua direita e esquerda. As barras azuis (formato de I) representam locais de inserção de unidades de repetição, onde foram observadas reads com inserções de 4 (uma unidade de repetição) ou 12 (três unidades de repetição) bases, caracterizando os alelos 9 e 11, respectivamente. Os dois tipos de inserções encontradas estão indicadas nos quadros destacados em vermelho. 69

Figura 12. Estrutura de repetição do <i>locus</i> TPOx retirada do STRBase. Disponível em: https://strbase.nist.gov/seq_info.htm	70
Figura 13. Estrutura de repetição do <i>locus</i> SE33 retirada do STRBase. Disponível em: https://strbase.nist.gov/seq_info.htm	71
Figura 14. Fragmentos sequenciados nas duas direções <i>paired-end</i> . Na Read 1 a região de repetição está entre o início e fim do fragmento sequenciado, enquanto que na Read 2 gerada pelo sequenciamento da direção oposta à Read 1, a região de repetição não foi completamente englobada, com o fim do fragmento sequenciado estando dentro da região de repetição. Adaptado de (Warshauer et al. 2013).	72
Figura 15. Imagem de dois alelos 15 (15 e 15b) encontrados na amostra PIG0073 do <i>locus</i> D3S1358. O alelo 15 apresentou 24 leituras de cobertura, enquanto que o alelo 15b apresentou 23. Suas respectivas nomenclaturas seguindo recomendações da Sociedade Internacional de Genética Forense (ISFG) se encontram a frente de cada alelo.	78

Lista de Tabelas

Tabela 1. Classificação da amostra populacional da região de Ribeirão Preto e comparação com os dados do IBGE coletados durante o censo de 2010 (IBGE 2010).	41
Tabela 2. Informações gerais e probabilidade de aderência ao equilíbrio de Hardy-Weinberg (p_{EHW}) para cada um dos 51 marcadores do SNPforID 52-plex em amostra populacional da região de Ribeirão Preto, SP. Os valores em negrito indicam desvios significativos ($\alpha = 0,05$) em relação às expectativas do modelo. Os SNPs estão apresentados em ordem decrescente de Poder de Discriminação.....	44
Tabela 3. Informações gerais e probabilidade de aderência ao equilíbrio de Hardy-Weinberg (p_{EHW}) para cada marcador do conjunto IISNPs (92 SNPs) em amostra populacional da região de Ribeirão Preto, SP. Os valores em negrito indicam desvios significativos ($\alpha = 0,05$) em relação às expectativas do modelo. Os SNPs estão apresentados em ordem decrescente de Poder de Discriminação.....	47
Tabela 4. Probabilidade de <i>Match</i> (PM) combinada e Poder de Exclusão (PE) combinado para cada conjunto e união dos 139 SNPs.....	50
Tabela 5. Estimativas de ancestralidade para a amostra populacional de Ribeirão Preto, SP, com base na análise dos sistemas SNPforID e IISNPs.....	53
Tabela 6. Frequências alélicas e parâmetros forenses estimados em amostra populacional da região de Ribeirão Preto (SP), considerando os 13 STRs autossômicos presentes no sistema CODIS original.....	58
Tabela 7. Frequências alélicas e parâmetros forenses estimados em amostra populacional da região de Ribeirão Preto (SP), considerando 17 outros STRs autossômicos presentes no ensaio deste estudo	60
Tabela 8. Análise dos genótipos de cada um dos STRs autossômicos em amostras aparentadas.	63
Tabela 9. Comparação dos parâmetros forenses obtidos pelos conjuntos 52-plex e IISNPs. .	76
Tabela 10. Comparação dos parâmetros forenses para o sistema SNPforID 52-plex entre diferentes populações.	77
Tabela 11. Probabilidade de <i>Match</i> e Poder de Exclusão para diferentes conjuntos de marcadores genéticos em diversas amostras populacionais de caráter global.....	80
Tabela 12. Valores da contribuição ancestral de europeus, africanos e ameríndios para cada classificação étnica estabelecida pelo IBGE para a cidade de Ribeirão Preto (censo IBGE 2010), utilizados para o cálculo da estimativa total de ancestralidade baseada em STRs autossômicos.....	82

Sumário

1. Introdução.....	15
1.1 Identificação humana	15
1.2 Desenvolvimento da Genética Forense.....	16
1.3 Banco de Dados de Perfis Genéticos	17
1.4 Marcadores Genéticos.....	18
1.4.1 STRs	19
1.4.2 SNPs	21
1.5 Sequenciamento de Nova Geração	24
1.5.1 Sequenciamento de Nova Geração no âmbito forense	25
2. Justificativa	28
3. Hipótese	29
4. Objetivos	30
5. Materiais e Métodos	31
5.1 Coleta e Amostra populacional	31
5.2 Procedimentos Laboratoriais	31
5.2.1 Extração de DNA e quantificação	32
5.2.2 Preparo das amostras	32
5.2.3 Sequenciamento.....	33
5.3 Procedimentos para Processamento de Dados do Sequenciamento	34
5.3.1 SNPs	34
5.3.2 STR.....	38
6. Resultados	41
6.1 Amostra populacional	41
6.2 SNPs.....	41
6.2.1 Genotipagem.....	42
6.2.2 Parâmetros Forenses e equilíbrio.....	43
6.2.3 Análise de Ancestralidade	50
6.3 Marcadores de determinação do sexo	53
6.4 STR	54
7. Discussão	64
7.1 Genotipagem.....	64
7.1.1 Marcadores de determinação do sexo: Amelogenina, DYS391 e Y-INDEL M175	64

7.1.2	SNPs	66
7.1.3	STRs	67
7.2	Desvios ao equilíbrio de Hardy-Weinberg	73
7.2.1	SNPs	73
7.2.2	STRs	74
7.3	Parâmetros forenses	75
7.4	Ancestralidade	81
8.	Conclusão	84
	Referências	86
	Apêndice A – Protocolo de processamento computacional para genotipagem de STRs	96
	Apêndice B – Populações do SNPforID e Projeto <i>1000Genomes</i> utilizadas na inferência de ancestralidade	105
	Apêndice C – Tabelas de comparações de frequências alélicas	107
	Anexo A – Termo de Consentimento Livre e Esclarecido (TCLE)	125
	Anexo B - Termo de Consentimento Livre e Esclarecido para armazenamento de material biológico	127
	Anexo C – Protocolo de análise laboratorial (<i>HaloPlex</i>) para sequenciamento de nova geração	129
	Anexo D – Protocolo de processamento computacional de dados de sequenciamento de nova geração	137

1. Introdução

1.1 Identificação humana

Existem inúmeros métodos para realizar a identificação humana, como, por exemplo, técnicas de antropologia forense, odontologia forense e papiloscopia, que consistem em análises altamente eficazes e de baixo custo quando comparadas à análise de DNA. Porém, devido à complexidade das amostras, estado de conservação dos vestígios, ou até mesmo falta de material comparativo, essas metodologias muitas vezes não são passíveis de utilização, recorrendo-se então à análise de DNA (Alonso et al. 2005; Mendes-Junior 2012).

O exame de DNA é extremamente útil na resolução de crimes, principalmente os de maior complexidade, como identificação de pessoas em desastres em massa, identificação de pessoas desaparecidas, casos de assassinatos ou estupros (crimes hediondos). O material biológico do criminoso, encontrado em locais de crime ou até mesmo transferido para a vítima (via estupro ou outros tipos de agressões), é analisado e fornece perfis únicos de DNA. Estes perfis são posteriormente comparados com perfis de suspeitos ou perfis gerados em outras cenas de crime, auxiliando as investigações tanto em termos de identificação quanto no estabelecimento de ligação com outros crimes. Deste modo, o DNA proveniente do material biológico possui um grande potencial para a individualização do criminoso, fornecendo alternativas que nenhuma outra área forense oferece, possuindo capacidade de distinguir indivíduos e de identificá-los com elevado nível de confiabilidade (Bond 2007; Butler 2015; van Oorschot et al. 2010).

Desde sua introdução no âmbito criminal, a análise de DNA vem contribuindo não só para identificação de criminosos, pessoas desaparecidas e mortos, mas vem também auxiliando pessoas que foram encarceradas erroneamente a provarem sua inocência. Um projeto criado nos Estados Unidos, chamado *The Innocence Project*, já auxiliou 354 pessoas a provarem sua inocência através do DNA, incluindo 20 presos que estavam no corredor da

morte. Além disso foram identificados 152 criminosos que realmente haviam cometido os crimes em questão (<https://www.innocenceproject.org/> acessado em: 01/03/2018).

1.2 Desenvolvimento da Genética Forense

A genética forense se iniciou por volta de 1900 com o sistema de grupos sanguíneos ABO, descrito por Karl Landsteiner, que permitiu uma primeira separação e agrupamento de indivíduos. Porém, o sistema se tornou padrão para as aplicações forenses apenas por volta de 1931. Logo em seguida, outros grupos sanguíneos e proteínas séricas e eritrocitárias foram caracterizados, possibilitando a produção de perfis com maior capacidade de discriminação (Goodwin et al. 2011).

Entre os anos de 1960 e 1970, o desenvolvimento das enzimas de restrição, do sequenciamento de Sanger (Sanger et al. 1997) e do *Southern blotting* impulsionou a biologia molecular ao permitir a análise da variabilidade do DNA. Em 1984, o geneticista Alec Jeffreys, utilizando a técnica de *Southern blotting*, demonstrou o potencial da aplicação de *loci* de repetições *in tandem* (VNTR – *Variable Number of Tandem Repeats*), gerando o que se denominou DNA *fingerprinting*, onde se poderia diferenciar indivíduos examinando a variação no padrão de bandas contendo VNTRs (Jeffreys et al. 1985). Entretanto, a análise possuía algumas limitações: longo tempo de análise, difícil comparação entre resultados de diferentes laboratórios, necessidade de grande quantidade e alta qualidade de amostras de DNA e dificuldades estatísticas associadas ao tipo de dado gerado (Butler 2005; Goodwin et al. 2011).

Em 1983, a PCR (*Polymerase Chain Reaction*) surge como a grande revolução da biologia molecular e genética forense. Este procedimento permite amplificar regiões específicas do DNA, garantindo maior sensibilidade em análises até mesmo em situações onde o DNA encontra-se degradado, diminuindo o tempo e a quantidade de amostra necessários para a produção de um perfil (Goodwin et al. 2011). A técnica de PCR passou a ser empregada também para análise de DNA mitocondrial (mtDNA) relacionada à identificação humana, devido ao sucesso obtido em análise de amostras degradadas e de baixa quantidade de DNA (como exemplo, amostras de cabelo) (Borsting and Morling 2015;

Higuchi et al. 1988). Logo em seguida, nos anos de 1990, surgem as primeiras análises de repetições curtas *in tandem* (STRs – *Short Tandem Repeats*), marcadores genéticos mais utilizados atualmente em análises forenses (Hagelberg et al. 1991; Jeffreys et al. 1992).

As análises forenses de DNA se expandiram a partir de 1995, com a criação do Banco de Dados Nacional de Perfis Genéticos do Reino Unido e, em 1998, o dos Estados Unidos. Os Bancos de Dados de Perfis Genéticos são formados, em sua grande maioria, por perfis de STRs presentes no sistema CODIS (*Combined DNA Index System*) proposto no final de 1997 pelo FBI (*Federal Bureau of Investigation*). Anteriormente composto por 13 STRs (CSF1PO; D3S1358; D5S818; D7S820; D8S1179; D13S317; D16S539; D18S51; D21S11; FGA; TH01; TPOx; vWA), passou a contar com mais sete STRs (D1S1656, D2S1338, D2S2441, D10S1248, D12S391, D19S433 e D221045) a partir de 2017, totalizando 20 STRs. Os bancos de dados de DNA destes países cresceram rapidamente e o surgimento de novos kits de STRs com maior número de *loci* e uso do Y-STRs, impulsionou ainda mais a genética forense (Butler 2015).

Com o passar dos anos, os procedimentos relacionados a análises de DNA se desenvolveram de forma acelerada em todos os aspectos, desde a extração e quantificação até equipamentos de detecção e sequenciamento. Acompanhando o progresso das técnicas, houve também rigoroso processo de padronização e controle de qualidade, buscando tornar a análise forense de DNA uma análise cada vez mais robusta e confiável (Goodwin et al. 2011; Lee et al. 2014).

Atualmente, as análises de DNA forense passam por um período de sofisticação, com desenvolvimento de *softwares* para análises de evidências mais complexas e aplicação de métodos de maior sensibilidade, o que vem fazendo com que se consiga retirar mais informações das análises de sequências, complementando os resultados tradicionalmente obtidos (Butler 2015).

1.3 Banco de Dados de Perfis Genéticos

A criação do banco de dados de perfis genéticos na Inglaterra e Estados Unidos foram os propulsores para expansão destes em outros países. O banco de dados de perfis genéticos é um importante recurso investigativo, auxiliando não somente em resolução de crimes, mas também auxiliando na busca e identificação de pessoas desaparecidas (Jakovski et al. 2017). No Brasil, os primeiros avanços para a criação destes bancos de dados surgiram apenas em maio de 2009 com uma parceria entre Polícia Federal e FBI (*Federal Bureau of Investigation*), que resultou na implementação do sistema CODIS no país, o qual inclui um *software* criado pelo FBI que interliga os bancos de dados mantidos em todo o território nacional, isto é, a Rede Integrada de Bancos de Perfis Genéticos (RIBPG), o que permite com que a Polícia Federal gerencie o Banco Nacional de Perfis Genéticos (Mendes-Junior 2012). Em 28 de maio de 2012, foi sancionada a lei 12.654 (http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/112654.htm), regulamentando a coleta e armazenamento do material biológico, e também relacionando os crimes determinantes para que os perfis de DNA dos criminosos sejam adicionados ao banco de dados. Relatórios públicos contendo informações estatísticas e de trabalhos realizados pelo RIBPG no Brasil, são gerados a cada seis meses (geralmente nos meses de maio e novembro). O relatório divulgado pelo Ministério da Justiça, com dados até novembro de 2017, nos mostra que 19 laboratórios estaduais e 1 laboratório da polícia federal já participam efetivamente na RIBPG, com 8.225 amostras já cadastradas criminalmente, sendo sua maioria de vestígios encontrados em cenas de crime, e 2.544 amostras relacionadas a pessoas desaparecidas. Até o momento, 376 perfis obtiveram coincidência auxiliando em 436 investigações (Ministério da Justiça e Cidadania 2017). O primeiro caso no país de coincidência de um perfil genético obtido de um identificado criminalmente com um perfil genético de um vestígio obtido de cena de crime ocorreu em janeiro de 2015 (Ministério da Justiça e Cidadania 2015).

1.4 Marcadores Genéticos

Marcadores genéticos são amplamente usados na identificação humana devido a sua capacidade de distinguir os indivíduos com alta acurácia. Dentre os marcadores, se destacam os SNPs (*Single Nucleotide Polymorphisms*) e os STRs (*Short Tandem Repeats*), sendo este último mais utilizado na prática forense (Grover e Sharma 2014). Adicionalmente, o emprego de marcadores caracterizados por inserções/deleções de poucos pares de base, denominados

INDELs, passou a ser avaliado mais recentemente (Ferragut et al. 2016). Os marcadores geram perfis altamente informativos. Parâmetros estatísticos, denominados parâmetros forenses, retratam esta informatividade em termos de probabilidade. Dentre eles temos a Probabilidade de *Match*, Poder de Discriminação, e Poder de Exclusão. Laudos e artigos científicos utilizam os parâmetros forense para demonstrar a informatividade dos marcadores utilizados, testar a segurança/poder no uso do polimorfismo para os fins que se destina e proporcionar maior entendimento do leitor. A informatividade de cada tipo de marcador é bastante variável, devido à diferença no número de alelos existente em cada *locus*, bem como de suas frequências em diferentes populações. Marcadores bi-alélicos, como SNPs e INDELs, apresentam menor informatividade por *locus*, necessitando a análise de um maior número de *loci* para alcançar uma boa informatividade. Já os STRs apresentam maior variabilidade de alelos, o que causa o aumento da informatividade de cada *locus* (Butler 2005; Goodwin et al. 2011).

1.4.1 STRs

A identificação humana por DNA é convencionalmente realizada por meio da análise de STRs. Por mais de 15 anos, os perfis de DNA vêm sendo gerados através da amplificação do material genômico por PCR e detecção por eletroforese capilar (CE – *Capillary Electrophoresis*) baseado no tamanho dos fragmentos do marcador STR (Yang et al. 2014). Sua abundância (em torno de 700.000 distribuídos no genoma humano) e variabilidade no genoma humano favorecem o uso deste tipo de marcador (Gymrek et al. 2016). Ainda, devido à grande utilidade para a análise e interpretação de misturas de materiais biológicos, e como elevado número de banco de dados de perfis genéticos pelo mundo são de perfis gerados por STR, estes marcadores deverão permanecer ainda por longo tempo em análises forenses.

Os STRs são sequências constituídas por repetições de uma unidade de dois a seis nucleotídeos, arranjadas consecutivamente, gerando repetições *in tandem*. Essas unidades são específicas para cada *locus* e de variável número de repetição, gerando diferentes alelos. Para a utilização na prática forense, os STRs devem apresentar determinadas características: a) ser altamente discriminatório; b) possuir alelos facilmente identificáveis; c) análise replicável em

diferentes materiais genéticos; d) geração de poucos fragmentos inespecíficos durante o processo de amplificação; e) compatibilidade com amplificação multiplex (Mendes-Junior 2012; Goodwin et al. 2011). A obtenção de perfis de STR através do sistema multiplex é relativamente simples e muito difundida. Há vários kits disponíveis no mercado que englobam STR autossômicos e ligados aos cromossomos X e Y.

Os primeiros kits que englobaram todos os marcadores presentes no sistema CODIS (13 STR) a surgirem no mercado foram *Identifiler* (*Thermo Fisher Scientific*) e *PowerPlex 16* (*Promega*), os quais incluíam 15 STRs e uma inserção/deleção de 6 pb presente no gene da Amelogenina, voltada para a determinação do sexo. Kits mais novos passaram a englobar um número maior de STRs, projetados para atender as novas necessidades do sistema CODIS e normas europeias, que incluem 8 novos marcadores em suas análises, aumentando ainda mais o Poder de Discriminação em uma única reação. Os kits *GlobalFiler* (*Thermo Fisher Scientific*) e *PowerPlex Fusion System* (*Promega*) amplificam os *loci* CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, FGA, TH01, TPOX, vWA, DYS391 e Amelogenina. Adicionalmente, o *GlobalFiler* permite a amplificação de SE33, enquanto que o *PowerPlex Fusion* inclui os marcadores Penta D e Penta E (Butler et al. 2012).

A **Figura 1** mostra uma tabela com os kits disponíveis no mercado e quais *loci* cada kit possui, mostrando uma relação entre os kits disponíveis e o sistema CODIS.

Cromossomos	Locus	CODIS 13 (1997-2017)	CODIS20 (2017 - atualmente) ESS12	(2009 -atualmente) Interpol (ISSOL)	(2010 - atualmente)	PowerPlex 16®	PowerPlex 18D®	PowerPlex ES/ESX 16®	PowerPlex ES/ESX 17®	PowerPlex 21®	PowerPlex CS7®	PowerPlex Fusion®	Profiler Plus™	COfiler®	SGM Plus®	Sefiler Plus™	SinoFiler™	MiniFiler™	Identifiler®	NGM™	NGM Select™	GlobalFiler™
		Loci recomendados				Promega kits STR							Life Technologies (ABI) Kits STR									
1q	D1S1656																					
1q	F13B																					
2p	TPOX																					
2p	D2S441																					
2q	D2S1338																					
3p	D3S1358																					
4q	FGA																					
5q	CSF1PO																					
5q	D5S818																					
6p	F13A01																					
6q	D6S1043																					
6q	SE33																					
7q	D7S820																					
8p	LPL																					
8q	D8S1179																					
9p	Penta C																					
10q	D10S1248																					
11p	TH01																					
12p	D12S391																					
12p	vWA																					
13q	D13S317																					
15q	FESFPS																					
15q	Penta E																					
16q	D16S539																					
18q	D18S51																					
19q	D19S433																					
21q	D21S11																					
21q	Penta D																					
22q	D22S1045																					
Xp,Xq	Amelogenina																					
Yq	DYS391																					

Figura 1. Relação de kits para obtenção de perfis de STRs presentes no mercado. Em amarelo destaca-se pares de loci que estão no mesmo cromossomo. Os loci presentes em cada kit estão indicados pelo preenchimento da célula nas colunas pertinentes de cores cinzas. As células preenchidas pela cor preta, indicam os loci que devem ser analisados pelo sistema CODIS, ESS (conjuntos de STRs analisados na Europa) e Interpol (adaptado de Butler et al. 2012).

1.4.2 SNPs

SNPs são definidos como uma mutação de ponto, resultado de uma substituição de um nucleotídeo por outro. Essa mutação gera alelos distintos, com sua grande maioria apresentando apenas dois alelos (bi-alélicos). Devido sua natureza bi alélica, sua

informatividade é reduzida, fato este que contribui para sua menor utilização em identificação humana quando comparado aos STRs, sendo necessário a análise de 50 a 100 SNPs para se chegar a um Poder de Discriminação semelhante ao dos conjuntos de STRs utilizados. Estima-se, ainda, que haja cerca de aproximadamente 1,42 milhões de diferenças entre dois genomas devido aos SNPs (International SNP Map Working Group 2011; Goodwin et al. 2011).

Apesar de ser menos utilizados para a identificação humana, SNPs apresentam algumas vantagens em relação a outros marcadores: a) bons resultados em amostras degradadas; b) técnicas automatizadas permitem a geração de perfis com milhares de SNPs simultaneamente; c) baixa taxa de mutação; d) produto da PCR são pequenos, podendo possuir a partir de 33 pares de base (Butler 2005; Canturk et al. 2014).

Muitos SNPs se encontram em regiões codificantes do genoma, estando relacionados a doenças e outros fenótipos. No âmbito forense, pode-se dividir a utilização dos SNPs em quatro áreas principais: 1) SNPs informativos de linhagem – presentes no mtDNA (DNA mitocondrial) e cromossomos Y, estes marcadores são utilizados em casos de pessoas desaparecidas e desastres em massa para relacionar parentescos ou paternidade, complementando também análises de STR; 2) SNPs informativos de ancestralidade – são chamados de AIMs (do inglês *Ancestry Informative Markers*), marcadores informativos de ancestralidade, possuindo diferentes frequências nas populações ao redor do mundo; 3) SNPs informativos de fenótipos – usados na inferência de características físicas de indivíduos (cor da pele, cor dos olhos, cor do cabelo), esta área vem apresentando grande potencial para auxílio na resolução de crimes; 4) SNPs para identificação – usados para identificação, fornecendo informações que possibilitam a individualização ou a exclusão de pessoas que não poderiam ser fonte de uma determinada amostra (Budowle e van Daal 2008; Canturk et al. 2014). Muitos estudos são voltados para a seleção de conjuntos de SNPs para identificação, onde destaca-se um conjunto de 52 SNPs gerado pelos esforços do projeto *SNPforID* (Sanchez et al. 2006) e outro de 92 SNPs designado como *IISNPs* (Pakstis et al. 2010).

1.4.2.1 SNPforID

O consórcio SNPforID se iniciou em 2003 com objetivo de estudar *loci* de SNPs para utilização em identificação humana com foco principal em análises forenses, porém englobando também análise de linhagens e predição de origem geográfica (Amigo et al. 2008).

Sanchez e colaboradores (Sanchez et al. 2006) analisaram os 52 SNPs do SNPforID em populações asiáticas e europeias, gerando uma Probabilidade de *Match* de $5,0 \times 10^{-19}$ e $5,0 \times 10^{-21}$ respectivamente, demonstrando o elevado Poder de Discriminação alcançado por este conjunto. Apesar do SNPs presentes não serem voltados para predição de origem populacional, as análises de ancestralidades apresentaram resultados consistentes com o conhecimento biogeográfico de cada população. Em outro estudo, observou-se que a análise feita em amostras degradadas e material biológico escasso apresentaram resultados melhores que os gerado por STRs em 21 de 26 casos analisados, tornando o ensaio uma alternativa possível para contribuir na identificação (Musgrave-Brown et al. 2007).

Pesquisadores de vários países em diferentes continentes já realizaram trabalhos observando o comportamento deste conjunto de SNPs em suas populações como Itália (Barbaro et al. 2012), Timor Leste (Santos et al. 2011), China (Wei et al. 2014). Na América do Sul, um estudo realizado no Chile relata que os parâmetros forenses obtidos de 50 SNPs são suficientes para a identificação forense rotineira do país, além de realizar uma análise da frequência de cada um dos 52 SNPs presentes no 52-plex relacionada a latitude do país, comparando as amostras de diferentes regiões (Norte, Sul e Central) (Moreno et al. 2014).

Utilizando a tecnologia de sequenciamento de nova geração *Ion Torrent* (Life Technologies), um estudo analisou três conjuntos de SNPs voltados para a ancestralidade (*Europlex*, 34-plex e *Pacifiplex*) juntamente com o conjunto 52-plex, obtendo genótipos com alta qualidade mesmo com quantidade de material biológico inferior a 0,3 ng. O referido trabalho ainda comparou genótipos obtidos pelo *Ion Torrent* e genótipos obtidos pelo método *SNaPshot* (Thermo Fisher Scientific), onde 395 dos 408 genótipos (96,8%) obtidos de três amostras foram concordantes. Tal trabalho mostrou ser possível a obtenção de genótipos de alta qualidade com quantidades inferiores a 0,3 ng de material biológico (Daniel et al. 2015).

1.4.2.2 IISNPs

Este painel, assim como o *52-plex*, é um painel voltado para a identificação humana, de modo que possa ser utilizado com sucesso em qualquer país do mundo. O painel consiste em ferramenta complementar em casos forenses, como em casos de DNA degradado, ou em que os marcadores padrões (STR) não apresentem bom desempenho (Pakstis et al. 2010).

O IISNPs se baseia na análise de 92 SNPs, tendo número suficiente de SNPs para que a informatividade seja pelo menos comparável com os valores apresentados pelos STRs do sistema CODIS. Os SNPs foram escolhidos dentre um conjunto de mais de 500 possibilidades, de acordo com as seguintes regras: apresentam elevada diversidade intra-populacional (heterozigose $> 0,4$) em populações humanas distribuídas ao redor do mundo, e associado a pouca variação na frequência ($F_{ST} < 0,06$) entre estas populações testadas (Kidd et al. 2006; Pakstis et al. 2010).

Este conjunto de SNPs vem sendo utilizado em estudos de sequenciamento de nova geração, analisados em conjunto com outros SNPs de identificação (Borsting et al. 2014; Zhang et al. 2017).

1.5 Sequenciamento de Nova Geração

Com surgimento do Projeto Genoma Humano, que se beneficiaria de maior rapidez e reduções drásticas no custo no sequenciamento, a demanda por novas tecnologias se tornava evidente. Em 2004, mesmo ano em que foi declarado a conclusão do sequenciamento de um genoma inteiro pelo método de Sanger, o Instituto Nacional de Pesquisa do Genoma Humano (do inglês *National Human Genome Research Institute - NHGRI*), deu início a um programa de financiamento à iniciativa de redução do custo de sequenciamento de genomas humanos. O crescente estímulo para o desenvolvimento e comercialização das tecnologias de sequenciamento mais robusto, fez com que surgisse em 2005 a primeira plataforma de sequenciamento de nova geração, baseada no método de pirosequenciamento, a *454 Life Sciences*. Desde então, vários outros equipamentos e tecnologias foram lançados, cada qual

apresentando suas especificidades para amplificação e detecção de sequências (Metzker 2010; van Dijk et al. 2014). Os equipamentos de sequenciamento de nova geração apresentam características que se diferenciam do método Sanger, onde ao invés de clonagem bacterianas de fragmentos de DNA ou amplificação de pequenos fragmentos por PCR, a técnica conta com um sistema de criação de bibliotecas de DNA, no qual milhões de reações de amplificação e sequenciamento são produzidas em paralelo, eliminando gasto de tempo e o intenso trabalho laboratorial de preparação das placas de sequenciamento, produção de clones ou separação dos fragmentos em géis (Anandhakumar et al. 2015).

A tecnologia de sequenciamento de nova geração continua em desenvolvimento. Uma plataforma desenvolvida pela *Pacific Biosciences*, a PacBio é uma tecnologia de terceira geração baseada em um instrumento que sequencia moléculas únicas e longas de DNA em tempo real, onde é detectado o nucleotídeo adicionado através do aumento da intensidade da fluorescência. Outra plataforma de alta tecnologia foi desenvolvida pela *Oxford Nanopore Technologies*, a Nanopore, a qual se destaca por seu baixo custo, alta velocidade de sequenciamento e a eliminação do procedimento de preparo de bibliotecas, onde a molécula de DNA ou RNA é sequenciada diretamente (Glenn 2011; van Dijk et al. 2014). Porém, ambas apresentam altas taxas de erros quando comparadas com plataformas NGS mais tradicionais como *Illumina* e *Ion Torrent* (Quail et al. 2012; Steinbock e Radenovic 2015).

Hoje, as plataformas de NGS mais confiáveis envolvem a tecnologia *Illumina*, apresentando excelente qualidade na geração de perfis, bem como um menor índice de erros comparado a outras plataformas (Goodwin et al. 2016).

1.5.1 Sequenciamento de Nova Geração no âmbito forense

A utilização de NGS em investigações forenses surge trazendo numerosos benefícios com diferentes possibilidades de análise. Com início das aplicações voltadas principalmente para identificação humana e determinação de traços fenotípicos, as plataformas de NGS tem a possibilidade de análises extremamente amplas, como análise de STRs, SNPs, INDELs genomas mitocondriais, cromossomos Y, ancestralidade, análises microbiológicas, análises de

DNA de plantas e animais, análises epigenéticas e de diferentes tipos de RNAs, e outras possibilidades (Børsting e Morling 2017; Butler 2015; Hunter 2018; Yang et al. 2014).

A análise de DNA forense mais frequentemente realizada nos laboratórios periciais da polícia, é geração de perfis de DNA através da técnica de PCR-multiplex juntamente com a eletroforese capilar (CE - *Capillary Electrophoresis*). Essa técnica, apesar de bem estabelecida, apresenta limitações quanto ao número de STRs analisados (kits comerciais atualmente contendo 24 STRs), problemas na análise de misturas biológicas e interferência de picos *stutter* (artefatos que surgem durante um erro da polimerase na amplificação por PCR, caracterizados principalmente por fragmentos amplificados com uma unidade de repetição a menos que o fragmento original) complicando a interpretação dos resultados (Børsting e Morling 2017).

Nos primeiros estágios de desenvolvimento da tecnologia de sequenciamento de nova geração, as *reads* geradas eram curtas (em geral variando de 25 a 100pb) não conseguindo abranger estruturas de repetição de maior comprimento e suas respectivas regiões flangeadora, o que limitava o número de STRs cujas *reads* pudessem ser alinhadas às *reads* do genoma de referência, impossibilitando a análise deste marcador. Adicionalmente, a alta taxa de erro de amplificação e/ou sequenciamento em áreas homopoliméricas ou repetitivas dificultava a identificação precisa dos alelos. Com o avanço das tecnologias as *reads* geradas tornaram-se maiores, o que possibilitou então análises mais acurada dos STRs (Børsting e Morling 2015; Van Neste et al. 2012; Yang et al. 2014).

Uma plataforma NGS não apresenta limitação de tipos e quantidade de marcadores. Este sequenciamento admite, simultaneamente, inúmeros marcadores voltados para diferentes análises. Pode-se sequenciar, por exemplo, SNPs de ancestralidade, SNPs de identificação humana, SNPs voltados para fenótipos e STRs, abrindo caminho para análises concomitantes de identificação humana, ancestralidade e características físicas de vários indivíduos. Adicionalmente, uma plataforma de NGS consegue em um único ensaio realizar a análise de diferentes marcadores em até 96 indivíduos ou mais, economizando tempo em casos de rápidas investigações (Børsting e Morling 2015).

A presença desses sequenciadores em laboratórios forenses, principalmente em países de terceiro mundo, depende ainda dos preços de equipamentos e kits para sequenciamento, além de desenvolvimento de *softwares* mais rápidos e efetivos para análise de elevada quantidade de dados em menor tempo. Algumas iniciativas já aparecem em destaque no campo acadêmico como os programas *STRait Razor* (Warshauer et al. 2013) e *toaSTR*

(Ganschow et al. 2017). Plataformas mais compactas de bancadas vem sendo desenvolvidas e podem ser uma alternativa promissora para área principalmente por sua flexibilidade (Borsting e Morling 2015; Børsting e Morling 2017; Yang et al. 2014).

Ainda, nos últimos anos houve um aumento de debates sobre o uso de NGS em genética forense. As vantagens do NGS em comparação a métodos mais tradicionais são muitas, porém um ponto importante que emerge com a aplicação das plataformas em análises forense é a questão ética. O armazenamento e quem terá acesso às informações geradas pelos perfis, aparece como ponto de questionamento devido a grande quantidade de informações geradas, podendo acabar revelando fatos indesejáveis, como predisposição a doenças, fatos esses que deverão ser tratados com cuidado e auxílio de legislações específicas (Hunter 2018; Yang et al. 2014).

2. Justificativa

A identificação humana por DNA aparece cada vez mais em tribunais, como prova crucial para elucidação de crimes, desastres em massa e identificação de pessoas. Neste contexto, vê-se um futuro promissor para a análise conjunta de SNPs e STRs por sequenciamento de nova geração (Wendt et al. 2016). Em cenas de crimes, o material biológico encontrado por vezes se apresenta em pequena quantidade e com níveis de degradação mais elevados, o que pode prejudicar a análise de STRs, não possibilitando a geração de perfis completos contendo todos os STRs analisados. Para isso os SNPs aparecem como uma alternativa para complementar as análises, proporcionando ganho de informações a partir de uma amostra (Gettings et al. 2015b; Phillips et al. 2008).

A análise de DNA em plataformas de sequenciamento de nova geração vem demonstrando grandes vantagens e inúmeras possibilidades de aplicação. Quando falamos de STRs, além da determinação de alelos em termos de unidades de repetição, abre-se a possibilidade de analisarmos sequências específicas de cada alelo de uma determinada amostra, o que gera um ganho de informatividade, além de abrir caminho para uma gama muito maior de tipos de materiais de análise (Borsting e Morling 2015; Børsting e Morling 2017; Yang et al. 2014)

Os conjuntos de marcadores (*52-plex*, *IISNPs* e STRs) utilizados neste trabalho foram inicialmente selecionados em estudos de populações principalmente europeias, africanas e asiáticas, com baixíssimos níveis de miscigenação (Kidd et al. 2006; Pakstis et al. 2010; Sanchez et al. 2006), com intuito de utilização global (Zeng et al. 2012). Apesar da existência de estudos em países de diferentes continentes, incluindo países da América do Sul, como Venezuela (Ruiz et al. 2012), Equador (Poulsen et al. 2011) e Chile (Moreno et al. 2014), nenhum estudo englobando todos estes marcadores foi feito Brasil, tampouco foram realizadas análises envolvendo sequenciamento de nova geração. Deste modo, se faz necessário a realização de trabalhos relacionados a estes conjuntos de marcadores, como forma de validação em uma população miscigenada como a brasileira, observando seu nível de informatividade e trazendo resultados para comparação e auxílio em análises futuras.

3. **Hipótese**

Conjuntos de marcadores genéticos voltados para identificação forense (SNPs e STRs) podem ser analisados simultaneamente por meio de metodologia *HaloPlex* de preparo de bibliotecas e sequenciamento de nova geração *Illumina*, gerando resultados de fácil interpretação e alta acurácia. Ainda, os conjuntos de SNPs apresentam grande informatividade forense e eficiência para análise de ancestralidade na população miscigenada brasileira, tornando passível sua utilização em casos reais.

4. Objetivos

Este estudo tem por objetivo principal analisar 140 SNPs, 30 STRs, uma INDEL no gene da Amelogenina e uma Y-INDEL em 340 indivíduos miscigenados da região de Ribeirão Preto (SP) utilizando a plataforma de sequenciamento de nova geração *MiSeq Personal Sequencer* (Illumina Inc.). Dois painéis de SNPs (SNPforID e IISNPs) e todos os STRs presentes nos kits *PowerPlex Fusion System* (Promega) e *GlobalFiler* (Thermo Fisher Scientific) serão analisados, a fim de:

- Avaliar se tais marcadores podem ser analisados com sucesso utilizando método *HaloPlex* na plataforma *Illumina* de sequenciamento de nova geração;
- Determinar, para cada marcador, as frequências alélicas, genotípicas e aderência ao equilíbrio de Hardy-Weinberg;
- Avaliar a informatividade forense (heterozigose esperada, Probabilidade de *Match*, Poder de Discriminação e Poder de Exclusão) de cada marcador e de cada conjunto de marcadores;
- Avaliar se os painéis de SNPs atingem a informatividade dos kits de STRs;
- Avaliar se os painéis de SNPs se mostram adequados para se estimar a ancestralidade genômica individual e populacional, apesar de não terem sido selecionados para esta finalidade.

5. Materiais e Métodos

O presente estudo foi aprovado em seus aspectos éticos pelo Comitê de Ética em Pesquisa da FFCLRP-USP, de acordo com o Processo CEP-FFCLRP CAAE n.º 25696413.7.0000.5407. Todos os participantes assinaram o Termo de Consentimento Livre e Esclarecido autorizando a utilização de seus materiais biológicos para o estudo (Anexo A e B).

5.1 Coleta e Amostra populacional

Foram coletados, com auxílio da equipe de enfermagem do Hemocentro de Ribeirão Preto, cerca de 10ml do material biológico (sangue periférico), em tubos contendo anticoagulantes (EDTA), sendo posteriormente armazenado a 4°C até o momento da extração. A amostra populacional foi constituída por 340 indivíduos (voluntários), sendo 167 mulheres e 173 homens com idades de 18 a 72 anos, da cidade de Ribeirão Preto e regiões vizinhas. Para cada participante coletou-se, ainda, dados referentes à classificação étnica (abordagem do IBGE) e ancestralidade por meio de preenchimento de questionário.

5.2 Procedimentos Laboratoriais

Todo procedimento experimental realizado pelo grupo de pesquisa do Laboratório de Pesquisas Forenses e Genômicas (LPFG) está descrito detalhadamente em protocolo presente no Anexo C desta dissertação. Entretanto, um breve relato dos procedimentos será apresentado nesta seção.

5.2.1 Extração de DNA e quantificação

O DNA foi extraído utilizando o protocolo *salting-out* (Miller et al. 1988) e re-suspendido em água, sendo armazenado a -20°C (Anexo C – item 1). A quantificação do DNA extraído foi realizada utilizando os equipamentos *NanoDrop™ 1000 Spectrophotometer* (Thermo Fisher Scientific) e *Qubit® 2.0 Fluorometer* (Thermo Fisher Scientific). Para verificar o nível de degradação das amostras realizou-se também eletroforese em gel de agarose (Anexo C – item 2). Considerando o resultado da análise fluorimétrica, as amostras foram diluídas para uma concentração final de 5 ng/μL com volume final em média de 100 μL.

5.2.2 Preparo das amostras

As bibliotecas de fragmentos para sequenciamento foram geradas por meio de ensaio customizado implementado no kit *HaloPlex Target Enrichment System* (Agilent Technologies, Inc.) de acordo com protocolo utilizadas da *Agilent Technologies* para sequenciamento *Illumina* (*HaloPlex Target Enrichment System For Illumina Sequencing*, Agilent Technologies, 2013). A seguir estão listados resumidamente os passos para a produção das bibliotecas (Anexo C – itens 3-5). A Figura 2 ilustra de forma simples as etapas realizadas.

- 1) *Digestão do DNA genômico*: nesta primeira etapa, as amostras de DNA genômico passam pelo processo de digestão com oito pares diferentes de enzimas de restrição, criando uma biblioteca de fragmentos de restrição de DNA genômico (gDNA).
- 2) *Hibridação dos fragmentos de gDNA*: os fragmentos de DNA são hibridados com milhares de sondas biotiniladas *HaloPlex* (customizada em função das regiões alvo do projeto).
- 3) *Captura do DNA alvo em forma circular*: sondas circularizadas contendo DNA alvo, são então capturados por esferas magnéticas revestidas por

estreptavidina, possibilitando a separação destes fragmentos do restante do DNA genômico não capturados (não alvo).

- 4) *Reação em cadeia da Polimerase (PCR)*: os fragmentos são amplificados por meio da reação de PCR (20 ciclos) no termociclador *SureCycler 8800* (Agilent Technologies).
- 5) *Purificação*: Após a amplificação, os fragmentos são purificados utilizando-se esferas magnéticas e re-suspendidos em Tris-HCL (pH 8.0).

Figura 2. Representação geral do preparo da biblioteca utilizando-se o kit *HaloPlex* (Adaptado de <https://www.agilent.com/cs/library/usermanuals/Public/G9900-90001.pdf>).

Para validação das bibliotecas empregou-se eletroforese capilar através do equipamento *2100 Bioanalyzer*, comparando resultados obtidos com perfil de espectro esperado (padrão) para as bibliotecas.

5.2.3 Sequenciamento

O sequenciamento foi realizado com quantidades equimolares (4nM) das bibliotecas combinadas em *pool* único (cada *pool* possuem até 96 bibliotecas) e diluído a 16pM de acordo com protocolo (Illumina 2013) para serem sequenciadas na plataforma *MiSeq Personal Sequencer* (Illumina Inc.). Para o sequenciamento utilizou-se os reagentes *MiSeq Reagent Kit V3 (600 cycle)* (Anexo C – item 6).

5.3 Procedimentos para Processamento de Dados do Sequenciamento

As metodologias descritas no item anterior, permitiram a análise simultânea de SNPs e STRs. Para a realização de genotipagem, foram seguidos processos distintos para cada tipo de marcador. Os procedimentos realizados para genotipagem de SNPs são os mesmos convencionalmente empregados pelo grupo de pesquisa do LPFG, e estão descritos detalhadamente em protocolo presente no Anexo D desta dissertação. Os procedimentos aqui estabelecidos para a genotipagem dos STRs foram descritos detalhadamente no Apêndice A.

5.3.1 SNPs

5.3.1.1 Trimagem e alinhamento

Os arquivos gerados pela plataforma *MiSeq* de sequenciamento de nova geração no formato FASTQ contendo as *reads* sequenciadas foram trimados (retirada de adaptadores inseridos no preparo da biblioteca) utilizando a ferramenta *cutadapt* (Martim 2011). Para o alinhamento das *reads* ao genoma humano de referência (GRCh37/Hg19) foi utilizado o algoritmo BWA-MEM (Li e Durbin 2009) contido no pacote BWA (*Burrows-Wheeler*) (Anexo D – itens 1.2.2).

5.3.1.2 Genotipagem

Após o mapeamento, três abordagens do pacote GATK (McKenna et al. 2010) são empregadas na rotina do LPFG para a obtenção dos genótipos de cada SNP (Anexo D – item 1.3.1): *HaplotypeCaller* (HC), *UnifiedGenotyper* (UG), e o *HaplotypeCaller* no modo GVCF (GVCF). Para genotipagem do conjunto de 52 SNPs (*52-plex*) foi utilizado duas das três abordagens, sendo elas HC e UG, o *HaplotypeCaller* é o método padrão indicado pelos desenvolvedores do GATK para realizar análises de SNPs, entretanto o UG apresenta melhor desempenho que o HC em situações de baixa cobertura e na genotipagem de SNPs tri-alélicos. Já para a genotipagem do conjunto de 92 SNPs foi utilizada apenas uma abordagem: o GVCF. Isso se deve pois, no decorrer da pesquisa a versão do UG foi descontinuada, não sendo mais atualizada por seus desenvolvedores. Paralelamente, houve a atualização da ferramenta HaplotypeCaller do pacote GATK, incluindo o modo mais completo de análise disponível até o momento, o GVCF, que passou então a ser o método recomendado para a finalidade de genotipagem de SNPs.

Após a genotipagem, os arquivos contendo os genótipos dos sítios de variação de interesse foram analisados com a ferramenta VCFx (Castelli 2015; Lima et al. 2016), que permite avaliar a qualidade dos genótipos identificados pelo GATK, interrogando aqueles que não se adequam a alguns parâmetros de cobertura e qualidade de sequenciamento pré-estabelecidos, como a qualidade das bases em cada *read*, o número de *reads* dando suporte ao genótipo e a qualidade da região flangeadora (Anexo D – item 1.3.2). No caso do *52-plex*, foi realizada então a comparação dos dados gerados pelas ferramentas HC e UG do pacote GATK, após passar por análise pelo VCFx. Pontos que apresentam diferenças ou genótipos inconclusivos, chamados de interrogados, eventualmente foram definidos após avaliação manual dos arquivos de alinhamento através do *software* IGV (*Integrative Genomics Viewer*) (Robinson et al. 2011; Thorvaldsdóttir et al. 2013).

5.3.1.3 Análises estatísticas

Para análises de frequências alélicas e genotípicas, heterozigose observada e esperada e análise de aderência ao equilíbrio de Hardy-Weinberg (HW) foi utilizado o *software* GENOPOP 4.51 (Rousset 2008).

Os parâmetros forenses (PM, PD e PE) foram calculados com auxílio da planilha *PowerStats v12 (Promega)* e do *software Microsoft Excel*.

- a) Probabilidade de *Match* (PM): é a Probabilidade deque dois indivíduos selecionados aleatoriamente na população possuam genótipos idênticos. O PM pode ser representado para um *locus* ou para vários *loci*, dessa forma quanto maior o número *loci* analisados, menor será a Probabilidade deque dois indivíduos apresentem mesmo perfil genotípico.

Cálculo para um *locus*:

$$P_M = \sum_{i=1}^n G_i^2$$

Onde G_i é a frequência de cada genótipo distinto; n é o número de genótipos distintos (Brenner e Morris 1990; Jones 1972).

Calculo para vários *loci*: é o produto de PM de cada *locus*.

- b) Poder de Discriminação (PD): é a Probabilidade deque dois indivíduos selecionados aleatoriamente na população possuam genótipos distintos. O PD pode ser representado para um *locus* ou para vários *loci*, dessa forma quanto maior o número *loci* analisados, maior será a capacidade de discriminar indivíduos de uma população.

Cálculo para um *locus*:

$$P_D = 1 - \sum_{i=1}^n G_i^2$$

Onde n é o número de genótipos distintos; G_i é a frequência de cada genótipo distinto (Brenner e Morris 1990; Jones 1972).

Cálculo para vários loci:

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

Onde P_{Di} é a Poder de Discriminação para cada *locus* e L é o número de *locus*.

- c) Poder de Exclusão (PE): é a probabilidade que um determinado *locus* apresenta de excluir um suposto pai que não seja o pai biológico em um exame de paternidade. O PE pode ser também representado para vários *loci* aumentando a possibilidade de uma possível exclusão em situações de paternidade ilegítima.

Cálculo para um locus:

$$P_E = H^2(1 - 2Hh^2)$$

Onde H é heterozigose observada e h é a homozigose observada (Brenner e Morris 1990).

Cálculo para vários loci:

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

Onde P_{El} é o Poder de Exclusão para cada *locus* e L é o número de *locus*.

5.3.1.4 Ancestralidade

As inferências de proporções de ancestralidades foram realizadas para ambos os conjuntos de SNPs (*52-plex* e *IISNPs*). Para inferência de proporções de ancestralidade individual e populacional utilizou-se o *software STRUCTURE* 2.3.4 (Hubisz et al. 2009; Pritchard et al. 2000) O modelo empregado foi o de mistura com frequências alélicas

correlacionadas; para isso os seguintes parâmetros foram utilizados: i) 100.000 passos de *burn-in*, seguido de 100.000 interações de *Markov Chain Monte Carlo*, ii) pré-definição de três clusters ($K=3$), onde cada um deles representaria uma das três ancestralidades principais da população brasileira, e iii) 100 corridas independentes. Para as análises do *52-plex*, foram utilizados dados genotípicos de populações parentais africanas, ameríndias e europeias (Apêndice B, Tabela B1), disponibilizados pelo consórcio SNPforID em <http://spsmart.cesga.es/snpforid.php> e dados de 14 populações da terceira fase do projeto *1000Genomes* (Apêndice B, Tabela B2). Para as análises com *IISNPs* apenas as mesmas 14 populações do projeto *1000Genomes* foram utilizadas. É importante mencionar que o projeto *1000Genomes* não inclui um grupo de ameríndios autóctones apropriados para representar a ancestralidade nativo-americana; deste modo, populações do leste asiático foram utilizadas a fim de substituir a população parental ameríndia em análises envolvendo populações deste consórcio.

A fim de comparação e confirmação dos resultados, estimativas adicionais de mistura populacional foram calculadas através de outros dois *softwares*: Admix95, abordagem *weighted* (Chakraborty 1985) e Admix 2.0 com abordagem *bootstrapped* (Dupanloup e Bertorelle 2001).

5.3.2 STR

5.3.2.1 Trimagem

Os arquivos gerados pela plataforma *MiSeq* de sequenciamento de nova geração no formato FASTQ contendo as *reads* sequenciadas foram trimados utilizando a ferramenta *cutadapt* (Martim 2011) (Anexo D – item 1.2.2).

5.3.2.2 Genotipagem

Foram utilizados dois programas para a realização da genotipagem dos STRs, que serão descritos brevemente a seguir:

a) *toaSTR* (Apêndice A - item 1.1.1)

O *toaSTR* é um *software* online de análise de STRs disponível em <https://toastr.de/>, sendo possível sua utilização em qualquer computador com acesso à internet. Para realização das análises deve-se criar uma conta, com *login* e senha para cada usuário, tendo acesso a todas as ferramentas disponíveis.

Foram analisados simultaneamente 26 marcadores incluindo 25 STRs (CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, DYS391, FGA, PENTA D, PENTA E, SE33, TPOx, TH01 e VWA) além da inserção/deleção de 6pb do gene da Amelogenina.

Os resultados gerados foram exportados em dois arquivos de saída de resultados: a) arquivos no formato .pdf contendo todas as informações resumidas, para cada marcador, de cada amostra; b) arquivos no formato .csv, para análises bioestatísticas, contendo apenas informações do *locus* e seus respectivos alelos, para cada amostra.

b) *Strait Razor* (Apêndice A – item 1.1.2)

Com base nos arquivos contendo as *reads* já trimadas, o *software STRait Razor v3* (Warshauer et al. 2015; Warshauer et al. 2013; Woerner et al. 2017) foi empregado para análise simultânea dos 30 STRs (CSF1PO, D1S1656, D2S441, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, DYS391, FES/FPS, FGA, F13A01, F13B, LPL, PENTA C, PENTA D,

PENTA E, SE33, TH01, TPOx e vWA) presentes no ensaio, além da inserção/deleção de 6pb do gene da Amelogenina.

Para cada amostra analisada, o *software STRait Razor* gera arquivos no formato .txt contendo uma lista dos *loci* de STRs analisados com a cobertura de cada alelo encontrado. Através de planilhas do programa *Microsoft Excel*, inclusas no pacote *STRait Razor*, o arquivo .txt é utilizado como *input* para determinar os genótipos (planilha *STRait Razor Analysis v3*) para todos os STRs de cada indivíduo. A confirmação visual dos genótipos determinados para cada STR pode ser feita pela observação das *reads* alinhadas ao genoma de referência hg19 utilizando-se o *software* IGV (Thorvaldsdóttir et al. 2013).

5.3.2.3 Análises estatísticas

Os dados obtidos por ambos os programas citados anteriormente (seção 5.3.2.2), foram organizados através de planilhas no *software Microsoft Excel* e comparados seus resultados, a fim de confirmar os genótipos obtidos entre os métodos. A determinação dos genótipos seguiu regras pré-estabelecidas descritas no Apêndice A (item 1.2). Após comparação de resultados, os parâmetros forenses (PM, PD e PE) foram calculados com auxílio da planilha *PowerStats v12 (Promega)* e do *software Microsoft Excel*.

6. Resultados

6.1 Amostra populacional

Os dados de classificação revelam que a distribuição da autoclassificação dos 340 voluntários participantes do presente estudo se dá de modo semelhante ao censo do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2010 (Tabela 1). Dentre estes 340 voluntários temos 172 homens e 168 mulheres.

Tabela 1. Classificação da amostra populacional da região de Ribeirão Preto e comparação com os dados do IBGE coletados durante o censo de 2010 (IBGE 2010).

	Presente estudo	IBGE		
		Ribeirão Preto	São Paulo	Brasil
Branco	64,12%	69,77%	63,92%	47,78%
Pardo	19,71%	22,84%	29,11%	43,12%
Preto	12,35%	6,37%	5,52%	7,61%
Amarelo	3,53%	0,93%	1,35%	1,10%
Indígena	0,29%	0,09%	0,10%	0,42%

6.2 SNPs

Todos os indivíduos foram submetidos aos processos de captura, montagem de biblioteca, sequenciamento e genotipagem com sucesso. Somando os SNPs presentes em ambos os conjuntos (SNPforID 52-plex e IISNPs), temos um total de 140 SNPs, pois estes apresentam em comum quatro SNPs: rs901398, rs2046361, rs1490413, rs891700. Destes 140, apenas um não foi possível ser analisado, o rs763869 presente no 52-plex.

6.2.1 Genotipagem

6.2.1.1 SNPforID - 52-Plex

A comparação entre os resultados gerados da genotipagem executada pelas ferramentas *HaplotypeCaller* (HC) e *UnifiedGenotyper* (UG) apresentou diferenças em 181 de um total de 17.340 genótipos analisados (1,04%) por ambas ferramentas. Tais diferenças foram manualmente verificadas no *software* IGV, considerando parâmetros como qualidade e quantidade de *reads*, incluindo o SNP e suas regiões flanqueadoras. Foi verificado que em 76,8% e 21% das diferenças o genótipo correto foi lido pelas ferramentas UG e HC, respectivamente; não foi possível definir a leitura correta em 2,2% dos 181 casos. Outros 169 genótipos (1%) foram determinados com base em baixa cobertura, de modo que um alelo foi dado como questionado pelos *softwares* GATK e/ou VCFx. Genótipos que foram dados como questionado não foram considerados em análises estatísticas. Desconsiderando-se estes dados com uma cobertura mais baixa, a cobertura média para cada marcador foi de 173 *reads* por indivíduo, com média de *reads* por marcador variando de 51 a 246 *reads*.

6.2.1.2 IISNPs

A genotipagem realizada pela ferramenta gVCF do *software* GATK, apresentou apenas 72 alelos interrogados. Os dados passaram pelo *software* VCFx para análise de qualidade aumentando para 238 (0,76%) os pontos interrogados dentre todos os genótipos gerados (31.280). Os 280 alelos interrogados foram verificados um a um no *software* IGV, restando então 111 genótipos questionados, que foram desconsiderados para as análises subsequentes. Grande parte dos genótipos excluídos se referem ao rs2920816 e apresentaram

baixa cobertura, com média de apenas 18 *reads*. A cobertura para cada SNP variou de 18 (rs2920816) até 706 (rs104357) *reads*.

6.2.2 Parâmetros Forenses e equilíbrio

6.2.2.1 SNPforID - 52-Plex

Os marcadores rs1015250 ($p = 0,0231$), rs2076848 ($p = 0,0369$), rs1528460 ($p = 0,0270$), rs2107612 ($p = 0,0309$) e rs1024116 ($p = 0,0202$) apresentam baixa probabilidade de aderência ao equilíbrio de Hardy-Weinberg (Tabela 2).

A informatividade de cada marcador foi analisada estatisticamente a partir dos parâmetros forenses: Probabilidade de *Match*, Poder de Discriminação e Poder de Exclusão (Tabela 2). A Probabilidade de *Match* variou de 0,3504 (rs1024116) a 0,5490 (rs938283), com média de $0,4036 \pm 0,0390$, enquanto que o Poder de Exclusão variou de 0,0550 (rs938283) a 0,2117 (rs2111980), atingindo média de $0,1475 \pm 0,0371$. A Probabilidade de *Match* combinada foi de $6,48 \times 10^{-21}$, enquanto que o Poder de Exclusão combinado foi de apenas 0,9997.

Tabela 2. Informações gerais e probabilidade de aderência ao equilíbrio de Hardy-Weinberg (p_{EHW}) para cada um dos 51 marcadores do SNPforID 52-plex em amostra populacional da região de Ribeirão Preto, SP. Os valores em negrito indicam desvios significativos ($\alpha = 0,05$) em relação às expectativas do modelo. Os SNPs estão apresentados em ordem decrescente de Poder de Discriminação.

<i>Continua</i>									
SNP	n^b	Cobertura média	Alelo referência (frequência) ^c	Heterozigose Observada	Heterozigose Esperada	Probabilidade de Match	Poder de Discriminação	Poder de Exclusão	p_{EHW}
rs1024116	340	184	C(0,5162)	0,4382	0,5618	0,3504	0,6496	0,1389	0,0202
rs1528460	340	150	C(0,4265)	0,4294	0,4899	0,3580	0,6420	0,1328	0,0270
rs1005533	340	215	G(0,4706)	0,4588	0,4990	0,3587	0,6413	0,1539	0,1675
rs914165	337	216	G(0,5074)	0,4645	0,5006	0,3592	0,6408	0,1583	0,2001
rs735155	339	173	C(0,5008)	0,4749	0,5006	0,3636	0,6364	0,1665	0,3849
rs2076848	337	182	A(0,6068)	0,4243	0,4779	0,3686	0,6314	0,1294	0,0369
rs1886510	338	197	G(0,5888)	0,4556	0,4850	0,3715	0,6285	0,1515	0,3019
rs8037429	339	114	C(0,5118)	0,4926	0,5005	0,3717	0,6283	0,1811	0,8283
rs1028528	338	214	A(0,6121)	0,4336	0,4756	0,3736	0,6264	0,1357	0,1056
rs1493232	340	203	C(0,3868)	0,4324	0,4751	0,3737	0,6263	0,1348	0,1122
rs2107612	283	51	G(0,3763)	0,4131	0,4699	0,3740	0,6260	0,1340	0,0309
rs354439	339	159	A(0,5251)	0,4956	0,4994	0,3741	0,6259	0,1837	0,9124
rs717302	335	131	G(0,4390)	0,4851	0,4933	0,3753	0,6247	0,1748	0,8227
rs1335873	339	235	T(0,4410)	0,4867	0,4938	0,3756	0,6244	0,1761	0,8241
rs1015250	338	182	G(0,6391)	0,4024	0,4620	0,3792	0,6208	0,1154	0,0231
rs1029047	332	130	T(0,6370)	0,4127	0,4631	0,3803	0,6197	0,1218	0,0576
rs727811	332	117	G(0,5225)	0,5105	0,4997	0,3814	0,6186	0,1969	0,7429
rs1454361	337	203	T(0,5875)	0,4866	0,4854	0,3839	0,6161	0,1761	1,0000
rs729172	339	221	G(0,6254)	0,4543	0,4693	0,3867	0,6133	0,1505	0,5535
rs891700 ^a	338	173	G(0,5104)	0,5235	0,4998	0,3878	0,6122	0,2089	0,4474
rs2831700	338	196	A(0,5340)	0,5237	0,4984	0,3900	0,6100	0,2091	0,3835
rs2111980	340	211	T(0,5191)	0,5265	0,5000	0,3900	0,6100	0,2117	0,3914
rs964681	340	195	T(0,6206)	0,4706	0,4706	0,3907	0,6093	0,1630	1,0000
rs737681	337	139	T(0,4080)	0,5015	0,4838	0,3927	0,6073	0,1888	0,5735
rs1490413 ^a	340	209	G(0,5941)	0,5000	0,4823	0,3927	0,6073	0,1875	0,5739
rs1355366	339	172	T(0,5841)	0,5074	0,4866	0,3929	0,6071	0,1940	0,4288
rs1463729	239	180	C(0,3746)	0,4720	0,4693	0,3936	0,6064	0,1641	1,0000
rs826472	307	62	T(0,3485)	0,4365	0,4549	0,3952	0,6048	0,1377	0,5333
rs1031825	336	116	A(0,3482)	0,4405	0,4546	0,3966	0,6034	0,1405	0,6326

									Conclusão
SNP	<i>n</i> ^b	Cobertura média	Alelo referência (frequência) ^c	Heterozigose Observada	Heterozigose Esperada	Probabilidade de Match	Poder de Discriminação	Poder de Exclusão	<i>p</i> EHW
rs2046361 ^a	336	185	A(0,6637)	0,5059	0,4812	0,3970	0,6030	0,1930	0,3691
rs722098	340	120	C(0,5809)	0,4167	0,4471	0,3973	0,6027	0,1244	0,2197
rs907100	340	225	T(0,6000)	0,5206	0,4876	0,3990	0,6010	0,2062	0,2195
rs251934	339	245	A(0,6563)	0,4513	0,4518	0,4031	0,5969	0,1483	1,0000
rs873196	339	177	C(0,3171)	0,4159	0,4337	0,4105	0,5895	0,1239	0,4540
rs1382387	340	202	C(0,3265)	0,4412	0,4404	0,4110	0,5890	0,1410	1,0000
rs917118	340	246	C(0,6294)	0,5059	0,4672	0,4115	0,5885	0,1927	0,1266
rs1979255	337	68	C(0,3299)	0,4527	0,4428	0,4126	0,5874	0,1493	0,7148
rs10495407	340	214	G(0,6941)	0,3941	0,4253	0,4142	0,5858	0,1104	0,2009
rs733164	339	231	G(0,6976)	0,4159	0,4225	0,4217	0,5783	0,1239	0,7944
rs719366	338	156	G(0,3092)	0,4527	0,4278	0,4275	0,5725	0,1493	0,3168
rs1360288	336	152	C(0,6994)	0,4405	0,4211	0,4301	0,5699	0,1405	0,4358
rs901398 ^a	337	370	G(0,2834)	0,4529	0,4247	0,4302	0,5698	0,1495	0,5121
rs2040411	337	134	C(0,3038)	0,3887	0,4068	0,4318	0,5682	0,1072	0,4160
rs876724	338	191	T(0,2737)	0,3876	0,3981	0,4402	0,5598	0,1065	0,6789
rs1413212	339	170	T(0,2802)	0,4248	0,4040	0,4425	0,5575	0,1297	0,4191
rs1357617	336	111	A(0,2530)	0,3750	0,3785	0,4580	0,5420	0,0994	0,8872
rs2830795	340	166	A(0,7500)	0,3647	0,3756	0,4598	0,5402	0,0939	0,6635
rs2056277	340	231	C(0,7603)	0,3441	0,3650	0,4690	0,5310	0,0834	0,2925
rs740910	339	183	A(0,7670)	0,3422	0,3580	0,4760	0,5240	0,0824	0,4453
rs2016276	339	228	T(0,7920)	0,3569	0,3299	0,5047	0,4953	0,0898	0,1430
rs938283	340	55	T(0,8250)	0,2790	0,2892	0,5490	0,4510	0,0550	0,5722

^a SNPs presentes em ambos os kits:52-plex e IISNPs;

^b Número de indivíduos genotipados com sucesso;

^c Frequência do alelo presente no genoma de referência hg19.

6.2.2.2 IISNPs

Os parâmetros forenses foram calculados igualmente para o conjunto *IISNPs*. A Probabilidade de *Match* variou de 0,3596 (rs10776839) a 0,4751 (rs10500617), com média de $0,3937 \pm 0,0236$, enquanto que o Poder de Exclusão variou de 0,1019 (rs8070085, rs12480506 e rs10500617) a 0,2538 (rs214955), atingindo média de $0,1675 \pm 0,0308$ (Tabela 3). A Probabilidade de *Match* combinada foi de $4,91 \times 10^{-38}$, enquanto que o Poder de Exclusão combinado foi de 0,99999997.

Apenas dois marcadores apresentaram baixa probabilidade de aderência ao equilíbrio de Hardy-Weinberg (Tabela 3): rs9866013 ($p = 0,0169$) e rs2175957 ($p = 0,0423$).

Tabela 3. Informações gerais e probabilidade de aderência ao equilíbrio de Hardy-Weinberg (p_{EHW}) para cada marcador do conjunto *IISNPs* (92 SNPs) em amostra populacional da região de Ribeirão Preto, SP. Os valores em negrito indicam desvios significativos ($\alpha = 0,05$) em relação às expectativas do modelo. Os SNPs estão apresentados em ordem decrescente de Poder de Discriminação.

Continua

SNP	n ^b	Cobertura média	Alelo referência (frequência) ^c	Heterozigose Observada	Heterozigose Esperada	Probabilidade de <i>Match</i>	Poder de Discriminação	Poder de Exclusão	p_{EHW}
rs10776839	340	333	G(0,528)	0,4618	0,4984	0,3596	0,6404	0,1562	0,1915
rs1294331	336	39	C(0,579)	0,4375	0,4875	0,3621	0,6379	0,1384	0,0581
rs7520386	340	471	G(0,426)	0,4471	0,4890	0,3635	0,6365	0,1452	0,1206
rs590162	329	32	C(0,520)	0,4742	0,4992	0,3639	0,6361	0,1659	0,3774
rs2269355	340	354	C(0,513)	0,4794	0,4997	0,3657	0,6343	0,1701	0,4491
rs321198	337	43	T(0,455)	0,4718	0,4960	0,3661	0,6339	0,1640	0,3803
rs6444724	340	111	T(0,550)	0,4706	0,4950	0,3666	0,6334	0,1630	0,3810
rs1821380	340	282	C(0,538)	0,4765	0,4971	0,3670	0,6330	0,1677	0,4462
rs1109037	340	321	G(0,447)	0,4706	0,4944	0,3672	0,6328	0,1630	0,3811
rs13182883	340	264	G(0,589)	0,4500	0,4838	0,3698	0,6302	0,1474	0,2173
rs1019029	340	323	A(0,501)	0,4912	0,5000	0,3707	0,6293	0,1799	0,7456
rs1554472	340	352	G(0,510)	0,4912	0,4998	0,3709	0,6291	0,1799	0,7456
rs10773760	339	221	A(0,580)	0,4631	0,4872	0,3713	0,6287	0,1572	0,3161
rs7205345	340	435	C(0,465)	0,4882	0,4976	0,3718	0,6282	0,1774	0,7440
rs987640	340	143	T(0,553)	0,4824	0,4944	0,3722	0,6278	0,1725	0,6614
rs338882	340	292	G(0,485)	0,4941	0,4996	0,3725	0,6275	0,1824	0,8288
rs10092491	340	208	T(0,428)	0,4735	0,4896	0,3732	0,6268	0,1654	0,5074
rs13218440	339	159	G(0,410)	0,4602	0,4838	0,3737	0,6263	0,1550	0,3703
rs521861	338	93	C(0,430)	0,4763	0,4902	0,3737	0,6263	0,1676	0,6571
rs6811238	340	193	T(0,490)	0,4971	0,4998	0,3738	0,6262	0,1849	0,9139
rs315791	340	272	A(0,494)	0,5000	0,4999	0,3751	0,6249	0,1875	1,0000
rs445251	339	178	G(0,403)	0,4572	0,4812	0,3753	0,6247	0,1527	0,3673
rs2811231	340	150	C(0,403)	0,4588	0,4812	0,3758	0,6242	0,1539	0,3691
rs1736442	340	154	T(0,376)	0,4235	0,4692	0,3761	0,6239	0,1289	0,0827
rs1498553	340	391	C(0,415)	0,4706	0,4856	0,3761	0,6239	0,1630	0,5773
rs4847034	340	327	A(0,412)	0,4706	0,4845	0,3772	0,6228	0,1630	0,6541
rs576261	340	178	A(0,557)	0,4912	0,4935	0,3773	0,6227	0,1799	0,9129
rs4796362	337	76	G(0,619)	0,4362	0,4717	0,3774	0,6226	0,1375	0,1669
rs1872575	340	118	G(0,472)	0,5029	0,4984	0,3780	0,6220	0,1901	0,9136

SNP	n ^b	Cobertura média	Alelo referência (frequência) ^c	Heterozigose Observada	Heterozigose Esperada	Probabilidade de <i>Match</i>	Poder de Discriminação	Poder de Exclusão	p _{EHW}
rs464663	340	271	T(0,397)	0,4647	0,4788	0,3804	0,6196	0,1584	0,5730
rs2503107	339	107	C(0,451)	0,5015	0,4952	0,3805	0,6195	0,1888	0,9126
rs722290	340	160	G(0,546)	0,5029	0,4958	0,3806	0,6194	0,1901	0,8275
rs7041158	340	109	C(0,616)	0,4500	0,4731	0,3807	0,6193	0,1474	0,3612
rs740598	340	276	G(0,432)	0,4941	0,4908	0,3813	0,6187	0,1824	1,0000
rs1523537	340	280	T(0,606)	0,4647	0,4775	0,3816	0,6184	0,1584	0,6497
rs2920816	289	18	A(0,621)	0,4464	0,4707	0,3818	0,6182	0,1447	0,3830
rs1358856	330	47	C(0,415)	0,4848	0,4856	0,3822	0,6178	0,1746	1,0000
rs3780962	340	137	A(0,499)	0,5147	0,5000	0,3827	0,6173	0,2007	0,6643
rs1410059	340	371	T(0,446)	0,5088	0,4942	0,3854	0,6146	0,1953	0,6606
rs10488710	340	263	C(0,403)	0,4824	0,4812	0,3855	0,6145	0,1725	1,0000
rs4364205	339	105	T(0,423)	0,4985	0,4881	0,3860	0,6140	0,1862	0,7394
rs12997453	340	172	A(0,400)	0,4824	0,4800	0,3866	0,6134	0,1725	1,0000
rs7229946	339	83	G(0,559)	0,5103	0,4930	0,3873	0,6127	0,1967	0,5819
rs891700 ^a	340	165	A(0,491)	0,5235	0,4998	0,3878	0,6122	0,2089	0,4474
rs2291395	340	192	A(0,612)	0,4765	0,4749	0,3890	0,6110	0,1677	1,0000
rs1336071	337	78	T(0,473)	0,5252	0,4985	0,3900	0,6100	0,2105	0,3820
rs2342747	340	353	A(0,353)	0,4294	0,4568	0,3904	0,6096	0,1328	0,2853
rs2272998	339	93	G(0,627)	0,4631	0,4677	0,3908	0,6092	0,1572	0,9075
rs4798798	340	153	C(0,618)	0,4765	0,4722	0,3917	0,6083	0,1677	0,8183
rs2399332	339	147	T(0,419)	0,5074	0,4869	0,3919	0,6081	0,1940	0,5032
rs2567608	339	170	T(0,489)	0,5310	0,4998	0,3921	0,6079	0,2161	0,2781
rs1490413 ^a	340	219	G(0,406)	0,5000	0,4823	0,3927	0,6073	0,1875	0,5739
rs6591147	340	276	T(0,387)	0,4853	0,4745	0,3936	0,6064	0,1750	0,7320
rs279844	337	61	A(0,543)	0,5282	0,4963	0,3940	0,6060	0,2134	0,2727
rs2046361 ^a	340	185	T(0,597)	0,5059	0,4812	0,3968	0,6032	0,1927	0,3691
rs5746846	340	615	C(0,451)	0,5324	0,4952	0,3975	0,6025	0,2174	0,1896
rs7704770	340	231	G(0,389)	0,4971	0,4758	0,3979	0,6021	0,1849	0,4936
rs430046	332	65	C(0,589)	0,4880	0,4838	0,3980	0,6020	0,1980	0,9055
rs13134862	340	209	G(0,647)	0,4529	0,4568	0,3980	0,6020	0,1495	0,5481
rs9905977	340	464	A(0,341)	0,4353	0,4494	0,3994	0,6006	0,1369	0,4875
rs159606	337	85	A(0,380)	0,4926	0,4712	0,4003	0,5997	0,1811	0,6411
rs993934	340	96	A(0,635)	0,4765	0,4636	0,4007	0,5993	0,1677	0,3075
rs689512	340	493	G(0,610)	0,5029	0,4758	0,4008	0,5992	0,1901	0,1039

SNP	n ^b	Cobertura média	Alelo referência (frequência) ^c	Heterozigose Observada	Heterozigose Esperada	Probabilidade de <i>Match</i>	Poder de Discriminação	Poder de Exclusão	p _{EHW}
rs985692	340	320	G(0,525)	0,5441	0,4988	0,4012	0,5988	0,2291	0,6385
rs4530059	340	427	G(0,638)	0,4765	0,4619	0,4023	0,5977	0,1677	0,1224
rs2073383	340	353	T(0,565)	0,5353	0,4916	0,4029	0,5971	0,2203	0,8115
rs221956	340	478	T(0,350)	0,4647	0,4550	0,4042	0,5958	0,1584	0,4143
rs1058083	340	440	A(0,368)	0,4882	0,4652	0,4044	0,5956	0,1774	0,2976
rs9606186	339	251	C(0,376)	0,4985	0,4692	0,4050	0,5950	0,1862	0,1434
rs3744163	340	279	G(0,594)	0,5235	0,4823	0,4053	0,5947	0,2089	0,6237
rs560681	340	296	A(0,671)	0,4294	0,4415	0,4054	0,5946	0,1328	0,1718
rs2292972	340	386	T(0,612)	0,5118	0,4749	0,4061	0,5939	0,1980	0,5754
rs9951171	340	242	G(0,596)	0,5118	0,4816	0,4061	0,5939	0,1980	0,1123
rs2255301	340	113	T(0,397)	0,5235	0,4788	0,4088	0,5912	0,2089	0,1658
rs1478829	340	111	T(0,378)	0,5088	0,4702	0,4093	0,5907	0,1953	0,0755
rs2175957	340	397	T(0,690)	0,3853	0,4278	0,4094	0,5906	0,1052	0,0423
rs8070085	340	278	A(0,693)	0,3794	0,4255	0,4107	0,5893	0,1019	0,7097
rs1027895	340	165	A(0,681)	0,4265	0,4345	0,4118	0,5882	0,1308	0,2356
rs2833736	340	151	G(0,647)	0,4882	0,4568	0,4126	0,5874	0,1774	0,2012
rs9866013	340	217	T(0,694)	0,3941	0,4247	0,4142	0,5858	0,1104	0,0169
rs214955	340	307	C(0,499)	0,5676	0,5000	0,4157	0,5843	0,2538	1,0000
rs8078417	340	216	C(0,699)	0,4206	0,4208	0,4236	0,5764	0,1269	0,8971
rs2270529	340	174	T(0,296)	0,4206	0,4168	0,4283	0,5717	0,1269	1,0000
rs9546538	340	204	T(0,706)	0,4176	0,4151	0,4288	0,5712	0,1250	1,0000
rs4288409	340	520	A(0,291)	0,4118	0,4126	0,4298	0,5702	0,1212	0,2514
rs901398 ^a	340	370	C(0,306)	0,4529	0,4247	0,4302	0,5698	0,1495	0,5121
rs6955448	340	229	C(0,707)	0,4324	0,4143	0,4340	0,5660	0,1348	0,1067
rs4606077	340	357	T(0,281)	0,4382	0,4041	0,4459	0,5541	0,1389	0,4132
rs10768550	340	142	C(0,726)	0,4176	0,3978	0,4466	0,5534	0,1250	0,8846
rs12480506	340	161	A(0,751)	0,3794	0,3740	0,4630	0,5370	0,1019	0,1958
rs1004357	340	706	G(0,251)	0,4029	0,3760	0,4641	0,5359	0,1157	1,0000
rs10500617	340	175	T(0,763)	0,3794	0,5000	0,4751	0,5249	0,1019	0,4526

^a SNPs presentes em ambos os kits: 52-plex e IISNPs;

^b Número de indivíduos genotipados com sucesso;

^c Frequência do alelo presente no genoma de referência hg19.

6.2.2.3 52-plex + IISNPs

Quando unimos os dois conjuntos para os cálculos dos parâmetros forenses, a Probabilidade de *Match* combinada obtida é de $1,20 \times 10^{-56}$, enquanto que o Poder de Exclusão combinado alcança o valor de 0,999999999936.

Tabela 4. Probabilidade de *Match* (PM) combinada e Poder de Exclusão (PE) combinado para cada conjunto e união dos 139 SNPs.

Parâmetros	52-Plex	IISNPs	139 SNPs
PM combinada	$6,37 \times 10^{-21}$	$4,91 \times 10^{-38}$	$1,20 \times 10^{-56}$
PE combinado	0,999719652505	0,9999999969867	0,999999999936

6.2.3 Análise de Ancestralidade

Considerando a natureza tri-híbrida da população brasileira, dois *softwares* (Admix95 e Admix 2.0) foram empregados para estimar a ancestralidade europeia, africana e ameríndia em níveis populacionais, enquanto que o *software* *STRUCTURE* foi utilizado para estimar ancestralidades a níveis populacionais e individuais. As estimativas geradas para cada indivíduo utilizando cada conjunto de SNPs estão representadas na Figura 3 e Figura 4. Os valores de proporções de ancestralidade determinados pelos *softwares* *STRUCTURE*, Admix95 e Admix 2.0, para a amostra populacional do presente estudo, estão apresentados na Tabela 5.

Figura 3. Diagrama triangular gerado pelo *software STRUCTURE* para os conjuntos SNPforID 52-plex e //SNPs, representando as estimativas de ancestralidade geradas para cada indivíduo. Cada indivíduo é representado por um círculo colorido, sendo em amarelo os indivíduos da população de Ribeirão Preto - SP. Indivíduos das populações parentais estão mais concentrados nos vértices do triângulo: EUR – Europeus (vermelhos); AMR – Ameríndios (SNPforID)/Asiáticos (1000Genomes) (azuis); AFR – Africanos (verdes).

Figura 4. Gráfico retangular gerado pelo *software STRUCTURE* representando as estimativas de ancestralidade geradas para cada indivíduo. Cada indivíduo é representado por uma barra que pode apresentar até três cores, sendo que vermelho, verde e azul correspondem a ancestralidade europeia (EUR), africana (AFR) e ameríndia/asiática (AMR/EAS), respectivamente.

Tabela 5. Estimativas de ancestralidade para a amostra populacional de Ribeirão Preto, SP, com base na análise dos sistemas *SNPforID* e *IISNPs*.

Conjuntos	Softwares empregados	Europeu	Africano	*Ameríndio/EAS
52-plex	Admix95 ^{a,c}	0,738 ± 0,010	0,188 ± 0,006	0,073 ± 0,008
Parentais:	Admix 2.0 ^b	0,730 ± 0,021	0,218 ± 0,013	0,052 ± 0,018
SNPforID	<i>STRUCTURE</i>	0,780	0,176	0,043
*Ameríndio	Média	0,749 ± 0,026	0,194 ± 0,0215	0,056 ± 0,015
52-plex	Admix95 ^{a,d}	0,739 ± 0,001	0,212 ± 0,001	0,050 ± 0,001
Parentais:	Admix 2.0 ^b	0,740 ± 0,021	0,210 ± 0,012	0,050 ± 0,021
1000Genomes	<i>STRUCTURE</i>	0,718	0,215	0,06
*EAS	Média	0,732 ± 0,0122	0,212 ± 0,003	0,053 ± 0,006
IISNPs	Admix95 ^{a,e}	0,704 ± 0,008	0,231 ± 0,006	0,065 ± 0,006
Parentais:	Admix 2.0 ^b	0,738 ± 0,053	0,198 ± 0,049	0,064 ± 0,040
1000Genomes	<i>STRUCTURE</i>	0,272	0,257	0,471
*EAS	Média	0,571 ± 0,260	0,229 ± 0,030	0,200 ± 0,235
52-plex + IISNPs	Admix95 ^{a,f}	0,735 ± 0,000	0,211 ± 0,000	0,054 ± 0,000
Parentais:	Admix 2.0 ^b	0,745 ± 0,022	0,206 ± 0,012	0,049 ± 0,020
1000Genomes	<i>STRUCTURE</i>	0,734	0,200	0,066
*EAS	Média	0,738 ± 0,006	0,205 ± 0,005	0,056 ± 0,009

^a Abordagem *weighted*; ^b Abordagem *bootstrapped*; ^c $R^2 = 0,996640$; ^d $R^2 = 0,999761$; ^e $R^2 = 0,999992$; ^f $R^2 = 0,999999$

6.3 Marcadores de determinação do sexo

Os marcadores da Amelogenina e o DYS391 foram analisados pelo *STRaitRazor* e *toastr* juntamente com os 29 STRs. A Y-INDEL rs2032678 foi analisada pelo mesmo método de análise dos SNPs, utilizando o *HaplotypeCaller* no modo *GVCF*.

Não houve discrepância entre sexo anotado nos questionários respondidos pelos voluntários e o sexo inferido pelos três marcadores, isto é, amostras femininas apresentaram apenas fragmentos com a deleção (cromossomo X) na INDEL da Amelogenina, enquanto que homens apresentaram os dois tipos de fragmentos (inserção indicativa de cromossomo Y e deleção indicativa de cromossomo X) na INDEL da Amelogenina, fragmentos indicativos de um único alelo no STR DYS391 e fragmentos indicativos da inserção ou deleção na Y-INDEL rs2032678. A Y-INDEL exibiu frequência de 0,982 e 0,178 para os alelos de referência (TTCTC) e alternativo (deleção), respectivamente, apresentando cobertura média de 62 *reads*, variando de 10 a 555 *reads* em cada indivíduo do sexo masculino. Os dados referentes aos marcadores da

Amelogenina e DYS391 estão descritos na seção 6.4 juntamente com os STRs autossômicos.

6.4 STR

Todos as regiões dos 31 marcadores, sendo 30 STRs mais a inserção do gene da Amelogenina, foram sequenciadas com sucesso.

Com a adição dos 6 STRs (FES-FPS, F13A01, F13B, LPL, PENTA C e SE33) que não estavam inicialmente inseridos no conjunto de marcadores normalmente analisados pelo *software STRaitRazor* (versão v3), todos os 31 *loci* presentes no ensaio da pesquisa puderam ser analisados. No entanto o *software* não detectou 5 marcadores: D2S441, D12S391, PENTA D, PENTA E e SE33. Já a ferramenta *toaSTR* conseguiu detectar todos os 25 STR e a inserção/deleção do gene Amelogenina que são propostos no painel de análise de tal ferramenta (Figura 5)

Os genótipos gerados pelos dois *softwares*, *STRaitRazor* e *toaSTR*, foram comparados e unificados em uma genotipagem final. Nenhuma diferença significativa foi encontrada durante essa comparação. As diferenças encontradas foram apenas relacionadas a baixa cobertura, ou a não captura de um determinado alelo ou marcador em uma das ferramentas.

Dos 31 marcadores presentes, após genotipagem final, oito (Amelogenina, CSF1PO, D16S539, D3S1358, D6S1043, D8S1179, TH01 e TPOx) foram analisados em 100% das amostras. O STR SE33 aparece com a menor porcentagem de genotipagem final alcançada com apenas 44%, ou seja, 148 amostras genotipadas.

Na Figura 6, podemos observar o sucesso de genotipagem em relação à quantidade de amostras. Das 340 amostras analisadas, 81 tiveram todos os seus STRs determinados com sucesso. Adicionalmente, apenas uma das amostras (PIG0130) teve menos de 50% de seus genótipos determinados.

Figura 5. Gráfico da porcentagem de sucesso de genotipagem dos 31 *loci* por cada *software* individualmente e após junção dos dados gerando genótipos finais.

Figura 6. Gráfico da percentagem de genótipos finais obtidos com sucesso e quantidade de amostras que alcançaram tal percentagem de sucesso.

As coberturas médias encontradas em cada *locus* variaram de 479 ± 308 (TPOx) a 21 ± 13 (D10S1284) no *software STRaitRazor* e de 485 ± 302 (TPOx) a $11 \pm 7,56$ (SE33) na ferramenta *toaSTR*, apresentando um alto desvio padrão além de elevada variação entre a cobertura dos *loci* (Figura 7).

Figura 7. Cobertura média de cada *locus* observada nos dois *softwares* empregados. As barras pretas representam o desvio padrão encontrado.

As Tabelas 8 e 9 mostram as frequências alélicas e parâmetros forenses dos 30 STRs presentes neste estudo. Na Tabela 6 estão os 13 STRs originalmente presentes no sistema CODIS, enquanto que na Tabela 7 está o restante dos 17 STRs presentes no ensaio. Entre os STRs autossômicos analisados na população de Ribeirão Preto, as frequências alélicas variaram de 0,0015 a 0,4493. Os valores de heterozigose observada variaram de 0,9025 (D1S1656) a 0,6526 (F13B). O STR SE33 apresentou o maior número de alelos diferentes com 33 alelos, já os *loci* F13B, LPL e TPOx tiveram uma menor variação, com apenas 6 alelos distintos. A Probabilidade de *Match* média foi de $0,0827 \pm 0,078$, variando de 0,0131 (SE33) a 0,1467 (LPL). O Poder de Exclusão variou de 0,2287 (LPL) a 0,8004 (D1S1656), com média de $0,5796 \pm 0,1374$.

Dos 29 STRs autossômicos analisados, 6 apresentaram desvios em relação ao equilíbrio de Hardy-Weinberg: D1S1656 ($p = 0,0033$), D2S1338 ($p = 0,0009$), D8S1179 ($p = 0,0429$), F13B ($p = 0,0000$), LPL ($p = 0,0000$) e SE33 ($p = 0,0000$).

Tabela 6. Frequências alélicas e parâmetros forenses estimados em amostra populacional da região de Ribeirão Preto (SP), considerando os 13 STRs autossômicos presentes no sistema CODIS original.

do sistema CODIS original.													Continua
Alelos	CSF1PO	D3S1358	D5S818	D7S820	D8S1179	D13S317	D16S539	D18S51	D21S11	FGA	THO1	TPOx	VWA
5												0,0015	
6											0,2118	0,0206	
7	0,0162		0,0118	0,0134			0,0015				0,2191	0,0059	
8	0,0206		0,0237	0,1607	0,0059	0,1250	0,0265				0,1235	0,4485	
9	0,0382		0,0340	0,1220	0,0162	0,0878	0,1382	0,0015			0,1794	0,1397	
9.3											0,2529		
10	0,2765		0,0577	0,2827	0,0824	0,0580	0,0956	0,0104			0,0132	0,0632	
11	0,3044		0,2944	0,2321	0,0838	0,3051	0,3029	0,0104				0,2721	0,0030
12	0,2824		0,3743	0,1503	0,1397	0,2619	0,2426	0,1376				0,0471	
13	0,0559	0,0029	0,1923	0,0357	0,2471	0,1190	0,1632	0,1213				0,0015	0,0060
14	0,0059	0,1062	0,0118	0,0030	0,2485	0,0417	0,0235	0,1450					0,0858
15		0,2758			0,1338	0,0015	0,0059	0,1243					0,1205
16		0,2552			0,0324			0,1361					0,2349
17		0,2109			0,0088			0,1346		0,0015			0,2726
18		0,1386			0,0015			0,0651		0,0137			0,1792
18.2										0,0091			
19		0,0088						0,0636		0,0775			0,0798
19.2										0,0015			
20		0,0015						0,0311		0,1292			0,0151
20.2										0,0061			
21								0,0104		0,1489			0,0030
21.2										0,0030			
22								0,0044		0,1505			
22.2										0,0061			
23										0,1353			
24								0,0030	0,0030	0,1383			
24.3										0,0015			
25								0,0015		0,1079			
25.2													
26									0,0015	0,0410			
26.2													
27									0,0346	0,0137			

													Conclusão
Alelos	CSF1PO	D3S1358	D5S818	D7S820	D8S1179	D13S317	D16S539	D18S51	D21S11	FGA	THO1	TPOx	VWA
27.1										0,0015			
28									0,1551	0,0091			
28.3									0,0045				
29									0,2229	0,0015			
29.3									0,0045				
30									0,2244				
30.2									0,0361				
31									0,0813				
31.2									0,0919	0,0030			
32									0,0030				
32.2									0,0798				
33									0,0015				
33.2									0,0361				
34									0,0060				
34.2									0,0030				
35									0,0105				
H _O	0,7147	0,8083	0,6893	0,8214	0,8029	0,7946	0,7912	0,8728	0,8343	0,8723	0,7618	0,7000	0,7771
H _E	0,7470	0,7849	0,7320	0,8026	0,8258	0,8258	0,7943	0,8881	0,8517	0,8824	0,7967	0,6996	0,8111
PD	0,8929	0,9157	0,8899	0,9287	0,9442	0,9296	0,9237	0,9727	0,9598	0,9716	0,9261	0,8666	0,9395
PM	0,1071	0,0843	0,1101	0,0713	0,0558	0,0704	0,0763	0,0273	0,0402	0,0284	0,0739	0,1334	0,0605
PE	0,4514	0,6145	0,4120	0,6394	0,6045	0,5891	0,5828	0,7402	0,6642	0,7393	0,5301	0,4283	0,5573
p _{EHW}	0,4373	0,0872	0,1722	0,6789	0,0429	0,3528	0,1963	0,1018	0,3062	0,2764	0,2289	0,4371	0,5973
2n	680	678	676	672	680	672	680	676	664	658	680	680	664

H_O: heterozigose observada; H_E: heterozigose esperada; PD: Poder de Discriminação; PM: Probabilidade de *Match*; PE: Poder de Exclusão; p_{EHW}: Probabilidade de aderência ao equilíbrio de Hardy-Weinberg; 2n: número de cromossomos considerados na análise de cada STR.

Tabela 7. Frequências alélicas e parâmetros forenses estimados em amostra populacional da região de Ribeirão Preto (SP), considerando 17 outros STRs autossômicos presentes no ensaio deste estudo

Continua

Alelos	D1S1656	D2S1338	D2S441	D6S1043	D10S1248	D12S391	D19S433	D22S1045	F13A01	F13B	FES-FPS	LPL	PENTA C	PENTA D	PENTA E	SE33	DYS391
2.2														0,0426			
3.2									0,1157								
4									0,0549								
5									0,1958				0,0148	0,0152	0,0978		
6									0,2522	0,1688	0,0051			0,0015			
7									0,3071	0,0308	0,0320	0,0023	0,0074	0,0106	0,1712		
8			0,0030						0,0193	0,2094	0,0034		0,0282	0,0441	0,0815		
9			0,0030		0,0015			0,0015		0,2175		0,0714	0,1558	0,1945	0,0245		0,0536
10	0,0061		0,2232	0,0103	0,0015		0,0044	0,0180		0,3701	0,2710	0,4493	0,0549	0,1246	0,0707	0,0068	0,5298
11	0,0549		0,3140	0,2345	0,0090		0,0222	0,1063	0,0045	0,0032	0,4158	0,2719	0,3635	0,1550	0,0897		0,4048
11.2																0,0068	
11.3			0,0610														
12	0,1006		0,0595	0,2581	0,0612		0,1080	0,0150	0,0030		0,2138	0,1705	0,2181	0,1474	0,1359		0,0119
12.2							0,0089									0,0068	
12.3			0,0030														
13	0,0732		0,0238	0,0826	0,2687		0,2456	0,0060	0,0134		0,0556	0,0346	0,1350	0,1702	0,1033	0,0270	
13.1			0,0015														
13.2							0,0251									0,0135	
13.3			0,0015														
14	0,1555		0,2693	0,0649	0,3284		0,2929	0,0509	0,0059		0,0034		0,0193	0,0714	0,0679	0,0507	
14.2							0,0503									0,0034	
14.3	0,0091													0,0015			
15	0,1646	0,0034	0,0357	0,0206	0,1821	0,0551	0,1435	0,3578	0,0208				0,0030	0,0198	0,0543	0,0473	
15.2							0,0488									0,0034	
15.3	0,0427																
16	0,1159	0,0738		0,0103	0,1254	0,0208	0,0311	0,3099	0,0045					0,0015	0,0380	0,0878	
16.2							0,0133									0,0068	
16.3	0,0518																
17	0,0381	0,2131	0,0015	0,0678	0,0194	0,1220	0,0030	0,1287	0,0030						0,0299	0,0743	
17.2							0,0030										
17.3	0,1204			0,0015		0,0149											
18	0,0061	0,0906		0,0973	0,0030	0,1935		0,0060							0,0082	0,1284	
18.2																	
18.3	0,0518					0,0238										0,0034	
19		0,1107		0,0885		0,1473									0,0163	0,0878	
19.1						0,0015											
19.2	0,0030															0,0034	
19.3	0,0030					0,0060											
20		0,1393		0,0428		0,1131									0,0054	0,0676	
20.1						0,0015											
20.2																0,0068	

Conclusão																	
Alelos	D1S1656	D2S1338	D2S441	D6S1043	D10S1248	D12S391	D19S433	D22S1045	F13A01	F13B	FES-FPS	LPL	PENTA C	PENTA D	PENTA E	SE33	DYS391
20.3	0,0030			0,0015		0,0030											
21		0,0419		0,0059		0,1012									0,0054	0,0270	
21.2																0,0135	
21.3				0,0103													
22		0,0705				0,0952										0,0203	
22.2																0,0203	
22.3				0,0015													
23		0,1258		0,0015		0,0551										0,0034	
23.2																0,0372	
24		0,0688				0,0357											
24.2																0,0169	
25		0,0503				0,0089											
25.2																0,0405	
26		0,0084				0,0015											
26.2																0,0338	
27		0,0034															
27.2																0,0338	
28.2																0,0507	
29.2																0,0236	
30.2																0,0270	
31.2																0,0135	
32.2																0,0068	
H _O	0,9024	0,7852	0,7768	0,8614	0,7612	0,8780	0,8373	0,7335	0,7626	0,6526	0,6667	0,5438	0,7715	0,8237	0,8913	0,8581	-
H _E	0,8948	0,8808	0,7710	0,8441	0,7681	0,8865	0,8156	0,7460	0,7874	0,7436	0,7049	0,6904	0,7745	0,8636	0,9039	0,9438	-
PD	0,9739	0,9713	0,9096	0,9556	0,9107	0,9742	0,9387	0,8944	0,9218	0,8908	0,8658	0,8533	0,9172	0,9634	0,9777	0,9869	0,5525
PM	0,0261	0,0287	0,0904	0,0444	0,0893	0,0258	0,0613	0,1056	0,0782	0,1092	0,1342	0,1467	0,0828	0,0366	0,0223	0,0131	0,4475
PE	0,8004	0,5719	0,5567	0,7174	0,5291	0,7507	0,6700	0,4820	0,5316	0,3588	0,3786	0,2288	0,5473	0,6438	0,7777	0,7109	-
pEHW	0.0033	0.0009	0.5086	0.7429	0.7310	0.5012	0.5132	0.1576	0.1483	0.0000	0.5842	0.0000	0.1674	0.0771	0.4606	0.0000	-
2n	656	596	672	678	670	672	676	668	674	616	594	434	674	658	368	296	336

H_O: heterozigose observada; H_E: heterozigose esperada; PD: Poder de Discriminação; PM: Probabilidade de *Match*; PE: Poder de Exclusão; pEHW: Probabilidade de aderência ao equilíbrio de Hardy-Weinberg; 2n: número de cromossomos considerados na análise de cada STR.

A Figura 8 mostra a taxa de ocorrência de *stutter* encontrada para cada *locus* ordenado de forma crescente. O *locus* FES-FPS apresentou a menor taxa com 0,00089 enquanto que o DYS391 apresentou a maior taxa com valor de 0,09457.

Figura 8. Gráfico da razão de *stutter* para cada *locus* de STR.

Dentre as amostras coletadas existem dois casos isolados de indivíduos aparentados mãe e filha. Essas amostras foram analisadas a fim de verificar se tais marcadores foram corretamente genotipados observando discordâncias entre os resultados obtidos. A Tabela 8 mostra o resultado da genotipagem em dois casos de perfis gerados pelos 29 STRs autossômicos analisados entre mãe e filha, onde o Caso I temos a Mãe I e Filha I e no Caso II temos a Mãe II e a Filha II. Houve diferença no Caso I apenas no *locus* D2S1338 enquanto que no Caso II observa-se diferenças nos *loci* LPL e FES/FPS.

Tabela 8. Análise dos genótipos de cada um dos STRs autossômicos em amostras aparentadas.

Marcadores	CASO I				CASO II			
	Mãe I		Filha I		Mãe II		Filha II	
	Genótipo		Genótipo		Genótipo		Genótipo	
AMEL	X	X	X	X	X	X	X	X
CSF1PO	12	12	10	12	10	11	10	11
D10S1248	13	16	13	15	14	14	14	16
D12S391	18.3	18.3	18.3	20	21	25	20	21
D13S317	8	11	11	12	8	11	11	11
D16S539	12	12	12	12	13	13	11	13
D18S51	14	21	14	15	12	16	13	16
D19S433	12	13	11	12	13	14	14	15
D1S1656	15	18.3	11	18.3	14	16	12	14
D21S11	28	31	31	31	28.3	29	28.3	32.2
D22S1045	15	16	15	16	16	16	16	16
D2S1338	18	18	22	22	17	17	17	17
D2S441	11	14	11	14	10	11.3	10	10
D3S1358	14	15	14	17	14	15	14	16
D5S818	11	13	13	13	12	12	12	12
D6S1043	17	18	18	20	19	18	12	18
D7S820	10	11	8	11	10	11	11	11
D8S1179	11	14	11	14	12	15	14	15
F13A01	5	15	4	15	6	7	7	7
F13B	10	10	9	10	6	8	8	11
FES/FPS	12	13	8	13	-	-	10	11
FGA	19	25	21	25	23	24	19	24
LPL	12	9	10	9	11	-	13	12
PENTA C	10	10	10	12	9	13	9	12
PENTA D	8	13	8	13	11	12	11	12
PENTA E	7	8	8	14	7	12	12	12
SE33	15	28.2	15	28.2	29.2	29.2	18	29.2
TH01	6	6	6	8	6	9.3	7	9.3
TPOx	8	11	8	11	8	10	8	11
vWA	14	17	17	18	15	18	15	18

*Alelos em cor vermelha representa alelo da filha que se originou da mãe (pares de alelo mãe/filhas); alelos em azul representam situações que não permitem determinar quais dos dois alelos foram transmitidos para a filha, pois mãe e filha possuem mesmo genótipo; destaques em negrito representam alelos originados do pai.

7. Discussão

7.1 Genotipagem

7.1.1 Marcadores de determinação do sexo: Amelogenina, DYS391 e Y-INDEL M175

A determinação do sexo (masculino e feminino) de amostras questionadas é de extrema importância principalmente em casos de abuso sexual, para a distinção e confirmação de material da vítima e do abusador (Butler 2005). Para isso, kits comerciais de STR passaram a incorporar polimorfismo presente no gene da Amelogenina. Este gene codifica a principal proteína presente na matriz do esmalte dos dentes e está presente tanto no cromossomo X quanto no cromossomo Y. O polimorfismo em questão consiste em uma INDEL, onde o alelo inserção está presente no cromossomo Y e o alelo deleção no cromossomo X. Mulheres apresentam duas cópias da deleção (XX), já homens apresentam uma cópia da deleção e outra da inserção (XY) (Lau et al. 1989; Srivastava et al. 2017). Nas Figuras 9 e 10 é possível visualizar, com auxílio do *software* IGV, a diferença encontrada entre homens e mulheres. Na Figura 9, ambas as amostras A (homem) e B (mulher) apresentam *reads* que alinham perfeitamente com a região do gene da Amelogenina presente no cromossomo X (AMELX). Na Figura 10, a amostra B não possui *reads* alinhadas na região do gene da Amelogenina presente no cromossomo Y (AMELY), enquanto que a amostra A apresenta grande quantidade de *reads* na região.

Figura 9. Imagem do *software* IGV indicando o alinhamento das *reads* na região do gene da Amelogenina do cromossomo X. Na parte superior (A) está representada amostra do sexo masculino; a parte inferior (B) representa amostra do sexo feminino.

Figura 10. Imagem do *software* IGV indicando o alinhamento das *reads* na região do gene da Amelogenina do cromossomo Y. Na parte superior (A) está representada amostra do sexo masculino; a parte inferior (B) representa amostra do sexo feminino.

A Amelogenina, assim como os STRs, é comumente analisada por técnicas de PCR e eletroforese capilar (Dutta et al. 2017; Mannucci et al. 1994; Srivastava et al. 2017). Uma rara deleção no gene da Amelogenina no braço *p* do cromossomo Y, pode causar a ausência do *amplicons* contendo a inserção de 6pb, proporcionando a falsa impressão de que uma amostra questionada masculina seja proveniente de uma mulher (Santos 1998). A chance de erros referentes à deleção no gene da Amelogenina pode variar dentre as regiões geográficas devido à diferente frequência de ocorrência da deleção entre as populações, tendo baixa ocorrência na Europa (Áustria com 0,02%) e maior ocorrência no sul da Ásia (Índia com 1,8%) (Kayser e Schneider 2009). A partir desta descoberta, os kits comerciais começaram a inserir marcadores exclusivamente encontrados no cromossomo Y a fim de contornar este problema. O STR DYS391 (braço *q* do cromossomo Y) e a Y-INDEL rs2032678 (braço *q* do cromossomo Y) estão presentes no kit *GlobalFiler*, enquanto que o kit *PowerPlex Fusion* apresenta apenas o STR DYS391.

Os resultados aqui obtidos apresentaram concordância de 100% em todas as análises envolvendo Amelogenina, DYS391 e Y-INDEL, determinando perfeitamente o sexo de cada voluntário. A amostra masculina PIG0130, por apresentar baixa cobertura associada a sequenciamento ruim, foi a única que mostrou disparidade nos três marcadores, captando apenas o cromossomo X na análise da Amelogenina e, ainda assim, com baixa cobertura. Vale ressaltar também que em mulheres, por terem dois cromossomos X, aparecem com coberturas mais elevadas quando comparados à cobertura encontrada para os homens no cromossomo X.

7.1.2 SNPs

A metodologia aqui utilizada permitiu a análise de 139 dos 140 SNPs propostos. A ferramenta *SureDesign* (Agilent Technologies, Inc.) não encontrou sondas capazes de capturar o segmento no qual se insere o SNP rs763869 presente no sistema *SNPforID 52-plex*, uma vez que a construção de tais sondas requer que a região de interesse seja flanqueada por sítios específicos para determinados pares de enzimas de restrição, fato este também relatado em trabalho anterior que já utilizava metodologia *Haloplex* (Wendt et al. 2016). Apesar desta limitação metodológica, que poderia ser solucionada pela amplificação separada deste SNP por *primers* específicos contendo às sequências adaptadoras necessárias para sequenciamento

Illumina (Lange et al. 2014), os resultados obtidos revelam alta eficiência de amplificação para os demais SNPs.

Este é o primeiro estudo que avalia a informatividade forense de marcadores dos sistemas *SNPforID 52-plex* e *IISNPs* em amostra de população miscigenada brasileira. Estudos de conjunto de SNPs analisados pela metodologia de minissequenciamento, empregando o conjunto de reagentes *SNaPshot Multiplex kit (Thermo Fisher Scientific)* eram normalmente utilizados (Sanchez et al. 2006), para a qual existem relatos de erros de genotipagem associados até mesmo a resultados que não apresentavam nenhuma dúvida (Lotte Hansen et al. 2007).

7.1.3 STRs

A genotipagem dos STRs se apresenta mais trabalhosa do que a dos SNPs. STRs necessitam de uma região sequenciada que englobe suas unidades de repetições e ainda regiões flanqueando cada lado das repetições. A fim de identificar os STRs, cada *software* utiliza segmentos conservados de ambas as regiões flanqueadoras de cada STR, contendo aproximadamente 12 nucleotídeos, que não necessariamente se encontram próximos às unidades de repetições, podendo estar até centenas de bases de distância. Uma vez que os segmentos utilizados pelos *softwares* para detecção dos STRs são distintos, essa variação de segmentos flanqueadores, além dos diferentes tipos de algoritmos utilizados, explica a diferença de cobertura e detecção dos marcadores entre os dois *softwares* empregados (Ganschow et al. 2017; Warshauer et al. 2013). Um ponto que vale destaque é a queda de qualidade nas bases incorporadas nos últimos ciclos do sequenciamento, comum no ensaio de sequenciamento *MiSeq* da *Illumina*. Deste modo, mesmo que a região flanqueadora desejada seja sequenciada, pode apresentar baixa qualidade principalmente em casos de STRs com *amplicons* maiores (alguns com 300pb), atrapalhando a análise (Erlich et al. 2008; Yang et al. 2013). Assim, o sucesso da genotipagem está diretamente relacionado ao tamanho dos alelos,

qualidade das *reads* e total cobertura das regiões flangeadoras que irão delimitar as regiões repetitivas dos marcadores.

Desta forma, a necessidade das *reads* englobarem todos esses elementos, faz com que os alelos maiores apresentem menor cobertura, como o *locus* FGA (médias próximas a 30 *reads*), o qual pode apresentar alelos contendo de 13 a 29 repetições, enquanto que STRs que apresentam alelos de tamanhos menores aparecem com uma cobertura mais elevada, como o *locus* TPOx (média próxima a 480 *reads*), cujos alelos podem apresentar de 4 a 14 repetições. Além disso, os elevados desvios apresentados nas médias de cobertura de cada *locus* de STR (Figura 7) também sofrem influência da grande variação de cobertura em cada amostra. As variações nos processos de extração de DNA genômico e de preparo de bibliotecas de fragmentos afetam a cobertura e, consequentemente, a genotipagem final.

Através do *software* IGV, conseguimos visualizar as unidades de repetição e verificar o sucesso do sequenciamento comparando com as *reads* geradas com a estrutura do STR em questão. Tomando o STR TPOx como exemplo, é possível visualizar claramente as unidades de repetição [AATG]_n (Figura 9) e comparar o padrão de *reads* gerado com a estrutura do STR TPOx descrita no STRBase (*Short Tandem Repeat DNA Internet DataBase*) (Figura 10) (STRBase). Adicionalmente, observa-se dois tipos de *reads*: algumas representando o alelo 9, as quais apresentam uma inserção de 4 bases se comparado ao alelo 8 do genoma de referência, e outras representando o alelo 11, as quais apresentam uma inserção de 12 bases quando comparado ao genoma de referência. Pode-se observar que toda a região flangeadora do *locus* TPOx foi sequenciada com sucesso, delimitando corretamente as regiões de repetição. Grande parte das *reads* sequenciadas contém ambas as regiões, flangeadora e de repetição, o que faz com que poucas *reads* sejam descartadas, garantindo elevada cobertura do *locus*.

Figura 11. Imagem do *software* IGV indicando a região do *locus* TPOx na amostra PIG0002, a qual possui genótipo 9/11. O genoma de referência (retângulo amarelo) apresenta alelo composto por 8 unidades de repetição. O quadro preto delimita a região repetitiva, e suas regiões flangeadoras à sua direita e esquerda. As barras azuis (formato de I) representam locais de inserção de unidades de repetição, onde foram observadas *reads* com inserções de 4 (uma unidade de repetição) ou 12 (três unidades de repetição) bases, caracterizando os alelos 9 e 11, respectivamente. Os dois tipos de inserções encontradas estão indicadas nos quadros destacados em vermelho.

Figura 12. Estrutura de repetição do *locus* TPOx retirada do STRBase. Disponível em: https://strbase.nist.gov/seq_info.htm

O *locus* SE33 foi o que apresentou menor cobertura dentre os demais com menor sucesso de genotipagem entre as amostras (148 amostras). A dificuldade apresentada por este *locus* já era esperada devido ao seu maior tamanho (3 a 39 unidades de repetição) e complexidade das regiões repetitivas: [AAAG]_n AG [AAAG]_nAG [AAAG]_n AAAAAG [AAAG]_n G AAGG [AAAG]_n AG (Figura 13). Adicionalmente, suas regiões flanqueadoras apresentam polimorfismos e regiões homopoliméricas, o que dificulta a interpretação do resultado do sequenciamento (Gettings et al. 2015a). A mesma dificuldade encontrada neste *locus* foi relatada em estudo anterior também realizado com a metodologia *HaloPlex* (Wendt et al. 2016).

Figura 13. Estrutura de repetição do *locus* SE33 retirada do STRBase. Disponível em: https://strbase.nist.gov/seq_info.htm.

O sequenciamento utilizando reagentes *MiSeq Reagent Kit V3*, gera *reads* do tipo *paired-end*, ou seja, o fragmento é sequenciado tanto na fita *forward* (*reads* R1) quanto na *reverse* (*reads* R2). Deste modo, além de alguns *amplicons* serem muito pequenos para captarem os STRs (excluídos no momento do desenho das sondas), outro problema encontrado no momento da genotipagem consiste na existência de alelos que aparecem em uma das fitas sequenciadas (R1 ou R2) mas não aparece na outra. Isso ocorre em situações onde uma região de repetição está situada no início de um fragmento *HaloPlex*, sendo o alelo detectado em uma das direções (*paired-end*), mas sequenciado parcialmente na segunda *read*, não englobando, deste modo as regiões flanqueadoras necessárias para análise de uma alelo dos STRs ou até mesmo parte da região repetitiva (Figura 14) (Warshauer et al. 2013). Nesses casos, genótipos homozigotos que foram detectados em apenas uma das fitas, foram desconsiderados nas análises. Vale ressaltar que o presente estudo visou apenas avaliar a executabilidade das análises em sequenciamento de nova geração. Assim sendo, a quantidade de *reads* médias previstas nas coberturas, podem não ser indicada para análise de casos reais. Ainda não foi estabelecido um valor mínimo de *reads* a ser considerado como aceitável.

Figura 14. Fragmentos sequenciados nas duas direções *paired-end*. Na *Read 1* a região de repetição está entre o início e fim do fragmento sequenciado, enquanto que na *Read 2* gerada pelo sequenciamento da direção oposta à *Read 1*, a região de repetição não foi completamente englobada, com o fim do fragmento sequenciado estando dentro da região de repetição. Adaptado de (Warshauer et al. 2013).

As análises de STRs apresentam frequentemente sequências com uma ou algumas unidades de repetição a menos que as de um alelo verdadeiro. Esses artefatos surgem no momento da PCR ocasionados por erros na atuação da DNA polimerase e são chamados de *stutter*. *Stutters* envolvendo mais de uma unidade de repetição de diferença podem existir, porém com rara ocorrência (Butler 2005). A presença de *stutter* é um dos grandes desafios na análise de STR, principalmente em casos de mistura de amostras com dois ou mais indivíduos, onde essas sequências podem confundir a real contribuição de cada indivíduo na amostra (Clayton et al. 1998). Estudos visando uma melhor interpretação e cálculo de *stutter* em sequenciamento de nova geração já começaram a ser realizados (Vilsen et al. 2018).

A presença de *stutter* pode variar de acordo com o tamanho do alelo, condições de amplificação e estrutura da sequência do *locus*. A taxa de ocorrência de *stutter* pode ser representada na forma de razão de *stutter* ou proporção de *stutter* (menos comum) (Vilsen et al. 2018; Zhang et al. 2018).

A Figura 8 mostra a razão de *stutter* para cada *locus*, sendo a razão da cobertura total do *stutter* pela cobertura total do alelo. Wendt et al, utilizando a metodologia *HaloPlex*, apresentou resultados semelhantes aos aqui obtidos, com uma presença maior de *stutter* nos *loci* DYS391, D18S51 e D12S391, enquanto que uma menor presença no *loci* PENTA D e TPOx (Wendt et al. 2016). Outro trabalho utilizando NGS para a análise de 34 STRs apresenta resultados consistentes com os obtidos por Wendt e o presente estudo (Zhang et al. 2018).

As análises, tanto de SNPs quanto dos STRs, entre indivíduos aparentados presentes no banco de dados das amostras coletas mostraram apenas três resultados discordantes entre mães e filhas. No Caso I apenas o *locus* D2S1338 apresentou diferença, provavelmente por não ter sido possível capturar o alelo 22 presente na mãe. Duas ressalvas aparecem no Caso II: (a) não foi possível genotipar o *locus* FES/FPS na mãe, o que é consistente com o fato de que este STR apresentou dificuldades em sua genotipagem, não sendo detectado em 43 das 340 amostras; (b) genótipo inconclusivo no *locus* LPL, o que decorre da baixa cobertura deste STR neste indivíduo, no qual apenas 4 *reads* foram detectadas. Deste modo possivelmente a mãe é heterozigota, sendo que o alelo transmitido para a filha não foi amplificado na mãe. As dificuldades de genotipagem deste *locus* se reflete no desvio do Equilíbrio de Hardy-Weinberg, justamente por deficiência de heterozigotos, e no baixo número (148, isto é, 43,5%) de amostras analisadas com sucesso.

7.2 Desvios ao equilíbrio de Hardy-Weinberg

Dizemos que uma população está em equilíbrio, quando as frequências alélicas e genotípicas não se alteram de uma geração para outra. Para tal, a população se enquadra em algumas premissas: tamanho da população muito grande, casamentos ocorrendo ao acaso, ausência de mutações e fluxo gênico, sem interferência de forças seletivas. Desvios ao equilíbrio podem indicar erros de genotipagem e amostragem (acaso), ou mesmo ação de forças evolutivas interferindo em determinado *locus* (Hartl e Clarke 2007).

7.2.1 SNPs

Os sete desvios encontrados em relação ao equilíbrio de Hardy-Weinberg (Tabelas 2 e 3), considerando ambos os conjuntos de SNPs (cinco no *52-plex* e dois no *IISNPs*), não

diferem do esperado. Devido ao nível de significância adotado ($\alpha = 0,05$), é estatisticamente esperado que seja observado aproximadamente 5% de desvios casuais. Neste caso, em 139 testes, seriam esperados 6,95 desvios. Vale ressaltar que dentre os sete marcadores, seis tiveram pequeno número de interrogações introduzidas pelo VCFx. Adicionalmente, a cobertura de sequenciamento destes marcadores foi elevada, variando de 150 a 397 *reads* em média por *locus*. Apenas o rs2107612 teve maior número de indivíduos excluídos (57) devido a genótipos interrogados, e uma média de cobertura menor que as demais com 51 *reads* presentes. Logo, a possibilidade de erros de genotipagem pode ser descartada. Além disso, se considerarmos a correção de Bonferroni para 139 testes ($\alpha_{\text{corrigido}} = 0,00035$), nenhum marcador apresentara desvios significativos em relação ao equilíbrio de Hardy-Weinberg.

7.2.2 STRs

Dos 29 STRs autossômicos analisados, o D1S1656 ($p = 0,0033$), D2S1338 ($p = 0,0009$), F13B ($p = 0,0000$), LPL ($p = 0,0000$) e SE33 ($p = 0,0000$) desviaram significativamente do equilíbrio de Hardy-Weinberg. Seguindo mesmo raciocínio feito para os SNPs, em 29 testes feitos, esperava-se apenas 1,45 desvios. Embora o número de cinco desvios observados difira significativamente dos 1,45 esperados ($\chi^2 = 9,1488$; $p = 0,0025$), se considerado o universo de 168 marcadores autossômicos analisados no presente estudo (29 STRs e 139 SNPs), o total de 11 desvios observados não difere dos 8,4 desvios esperados ($\chi^2 = 0,8471$; $p = 0,3574$). Deste modo, podemos descartar a ação de forças evolutivas sistemáticas, como, principalmente, mutação, migração deriva genética e seleção natural. Mesmo aplicando a correção de Bonferroni para 29 testes ($\alpha_{\text{corrigido}} = 0,00172$), ainda se observa quatro STRs (D2S1338, F13B, LPL e SE33) desviando do equilíbrio. Apesar disso, é provável que os problemas destacados no item 7.1.3, relacionados a genotipagem e número de amostras analisadas, possam explicar mais adequadamente os desvios ao equilíbrio de Hardy-Weinberg apresentados.

Os *loci* D2S1338, LPL, F13B e SE33 apresentam déficit de heterozigose, com diferenças entre heterozigose esperada e heterozigose observada de aproximadamente 10%, 15%, 10%, 11%, respectivamente. O D1S1656 não apresentou grande diferença de

heterozigose (excesso de 0,76%). Como sabemos, alelos maiores, podem não ser sequenciados completamente, gerando *amplicons* que não cobrem toda extensão de bases necessária para identificação destes alelos, não sendo detectados nas análises, proporcionando erros de genotipagem. Desta forma pode-se haver heterozigotos equivocadamente genotipados como homozigotos.

O Apêndice C apresenta tabelas de frequências alélicas de 28 dos 29 STRs autossômicos analisados no presente estudo, comparado os dados aqui obtidos com os descritos em outros estudos realizados no Brasil, principalmente na região Sudeste. Analisando a Tabela C3 (Apêndice C), podemos observar que o *locus* D2S1338 nas populações de Minas Gerais (MG) (del Castillo et al. 2009) e São Paulo (SP2) (Almeida Prado Oliveira e Sousa et al. 2014) possui os alelos 24, 25 e 26 (alelos de maior comprimento) com frequências próximas a 9%, 7,3% e acima de 1% respectivamente, enquanto que estes mesmo alelos foram sub-representados no presente estudo (6,9%, 5% e 0,8%, respectivamente). O *locus* SE33 (Tabela C25 do Apêndice C) também apresenta alelos maiores (com mais de 27 repetições) sub-representados em nossa amostragem, com diferença média de 2% em relação às frequências obtidas na população de São Paulo (SP2) (Almeida Prado Oliveira e Sousa et al. 2014). Estas diferenças entre frequências alélicas podem estar relacionadas à maior dificuldade de sequenciamento de alelos grandes mencionada no parágrafo anterior. Entretanto, podem estar simplesmente refletindo variações casuais entre frequências alélicas de diferentes amostras populacionais.

A baixa média de cobertura entre os marcadores citados (médias abaixo de 50 *reads*), indica dificuldade no sequenciamento dessas regiões, o que, da mesma forma, pode gerar erros na genotipagem, ocultando alelos que poderiam existir. O *locus* SE33 (11 *reads*) e o D2S1338 (12 *reads*) possuem as duas menores médias observadas.

Ao mesmo tempo, os marcadores LPL e SE33 apresentam baixo número de amostras analisadas, com 148 e 217 amostras, respectivamente, o que pode levar a erros de amostragem, não havendo uma representação correta da população. Este baixo tamanho amostral pode estar refletindo a dificuldade de definir genótipos homozigotos para alelos longos, ou heterozigotos envolvendo alelos longos.

7.3 Parâmetros forenses

Ambos os conjuntos de SNPs analisados se mostraram altamente informativos no que se refere a identificação individual (Probabilidade de *Match* combinada $6,48 \times 10^{-21}$ para *52-plex* e $4,91 \times 10^{-38}$ para *IISNPs*). Quando olhamos para a informatividade de vínculo genético (Poder de Exclusão combinado), o conjunto com 51 marcadores se mostrou menos informativo (PE combinado = 0,9997). Como esperado, devido ao maior número de marcadores o conjunto de 92 SNPs apresenta uma maior informatividade (PE combinado = 0,99999997).

Na Tabela 9 temos uma comparação entre os parâmetros forenses obtidos dos dois conjuntos. Foram selecionados três conjuntos de 51 SNPs, 51 mais informativos, 51 menos informativos e 51 selecionados aleatoriamente dentre os 92 SNPs presentes no conjunto *IISNPs* a fim de comparação com os 51 SNPs do conjunto *SNPforID*. A informatividade de vínculo genético foi a mesma para todos os três conjuntos selecionados (0,99993), sendo apenas 10 vezes maior que a do conjunto *52-plex*. Através dos valores de Probabilidade de *Match* calculados, observamos que a informatividade dentro do conjunto de 92 SNPs possui grande variação, onde o conjunto de 51 mais informativos do *IISNPs* se apresenta 24,5 vezes mais informativo do que os 51 SNPs do *52-plex*. Da mesma forma os 51 menos informativos do *IISNPs* foi menos do que duas vezes menos informativo que o *52-plex*. De modo geral, quando selecionamos aleatoriamente 51 SNPs conseguimos uma informatividade muito próxima da atingida pelos *52-plex*, mostrando que a diferença na informatividade entre os conjuntos *52-plex* e *IISNPs* é muito mais influenciada pelo número de marcadores presente em cada conjunto do que devido às características dos marcadores presentes em cada conjunto.

Tabela 9. Comparação dos parâmetros forenses obtidos pelos conjuntos *52-plex* e *IISNPs*.

Parâmetros	<i>52-plex</i>	<i>IISNPs</i>	*Menos Informativo	*Mais informativo	*Selecionados Aleatoriamente
PM combinado	$6,37 \times 10^{-21}$	$4,91 \times 10^{-38}$	$1,24 \times 10^{-20}$	$2,60 \times 10^{-22}$	$1,42 \times 10^{-21}$
PE combinado	0,9997	0,99999997	0,99993	0,99993	0,99993

*valores obtidos de 51 SNPs selecionados dentre os 92 SNPs presentes no conjunto *IISNPs*.

Apesar de realizada a análise de apenas 51 SNPs dos 52 SNPs presentes no ensaio 52-plex, a informatividade observada no presente estudo foi superior à de seis das sete populações apresentadas na Tabela 10, todas elas apresentando os 52 SNPs devidamente analisados.

Tabela 10. Comparação dos parâmetros forenses para o sistema SNPforID 52-plex entre diferentes populações.

População	Probabilidade de <i>Match</i> combinada	Poder de Exclusão combinado	Referência
África	$4,80 \times 10^{-17}$	0,9971	(Sanchez et al. 2006)
Europa	$5,00 \times 10^{-21}$	0,9998	(Sanchez et al. 2006)
Leste Asiático	$5,00 \times 10^{-19}$	0,9991	(Sanchez et al. 2006)
Oceania	$5,20 \times 10^{-16}$	0,9958	(Santos et al. 2011)
Timor Leste	$2,00 \times 10^{-18}$	0,9987	(Santos et al. 2011)
América^a	$2,50 \times 10^{-19}$	0,9993	(Moreno et al. 2014)
Chile	$4,90 \times 10^{-20}$	0,9999	(Moreno et al. 2014)
Ribeirão Preto	$6,48 \times 10^{-21}$	0,9997	Presente Estudo

^aPopulação SNPforID

Kits de INDELs (inserções-deleções) estão sendo desenvolvidos e apresentam características vantajosas para prática forense similares às dos SNPs, como a análise em pequenos *amplicons* e a baixa taxa de mutações. Estudos realizados com o conjunto de INDELs incorporados ao kit *Investigator DIPplex* contendo 30 INDELs bi-alélicos, revelaram parâmetros que atestam menor informatividade do que as inferidas para ambos os kits 52-plex e IISNPs, o que é consistente com o menor número de marcadores utilizados. Análises no Brasil (Ferreira Palha et al. 2015), Japão (Nunotani et al. 2015), Polônia e Taiwan (Pepinski et al. 2013) revelaram Probabilidade de *Match* de $3,4 \times 10^{-13}$, $2,67 \times 10^{-11}$, $7,98 \times 10^{-14}$ e $1,22 \times 10^{-11}$, respectivamente, valores inferiores em 6 ordens de grandeza em relação aos 52 SNPs e até 21 ordens de grandeza em relação aos 92 SNPs.

Os parâmetros forenses variam para cada *locus* de acordo com a quantidade de alelos que cada um apresenta. Os STRs, por apresentarem maior variedade de alelos, possuem uma maior informatividade quando comparados aos SNPs. O *locus* SE33 com 33 alelos, mesmo com todos os problemas de sequenciamento e genotipagem aqui já relatados, foi o que apresentou maior Poder de Discriminação (PD = 0,9869), o que o torna o mais informativo dentre todos os *loci*, incluindo STRs e SNPs, já que o SNP mais informativo apresenta Poder de Discriminação de apenas 0,6496 (rs1024116). No presente estudo, são necessários de 3 a 4 SNPs para atingir a média de informatividade de Poder de Discriminação alcançada por um

único STR, e de 4 a 6 SNPs para atingir a informatividade média de Poder de Exclusão atingida por um STR, dependendo do grau de informatividade de cada SNP.

Uma das grandes vantagens das tecnologias de NGS é a possibilidade de avaliar a variação de sequência de cada alelo, aumentando a informatividade. Quando se realiza eletroforese capilar, alelos com números iguais de repetições, porém com variações dentro de suas sequências, não conseguem ser distinguidos por apresentarem mesmos tamanhos (Borsting e Morling 2015; Van Neste et al. 2012). A Figura 15 ilustra dois alelos 15 encontrados pelo programa *toaSTR* no *locus* D3S1358, contendo variações nas suas unidades de repetições, sendo o alelo 15 contendo uma unidade (TCTA), uma unidade (TCTG) e 13 unidades (TCTA), enquanto que o alelo 15b contém uma unidade (TCTA), três unidades (TCTG) e 11 unidades (TCTA). Alguns estudos já vem sendo realizados verificando a variação de sequências em alelos de STR e mostrando um aumento da informatividade quando tal nível adicional de variabilidade é considerado nos cálculos forenses (Gelardi et al. 2014; Rockenbauer et al. 2014).

Porém, a Sociedade Internacional de Genética Forense (ISFG – *International Society for Forensic Genetics*) relata que os dados obtidos até hoje, são insuficientes para quantificar qualquer novo alelo de STR. São necessários mais estudos, de forma padronizada e com nomenclaturas adequadas, para comparação dos dados entre vários laboratórios (Parson et al. 2016).

15	A	24	0	D3S1358[CE15]-Chr03-GRCh38 45540739-45540802 [TCTA]1 [TCTG]1 [TCTA]13
15b	A	23	0	D3S1358[CE15]-Chr03-GRCh38 45540739-45540802 [TCTA]1 [TCTG]3 [TCTA]11

Figura 15. Imagem de dois alelos 15 (15 e 15b) encontrados na amostra PIG0073 do *locus* D3S1358. O alelo 15 apresentou 24 leituras de cobertura, enquanto que o alelo 15b apresentou 23. Suas respectivas nomenclaturas seguindo recomendações da Sociedade Internacional de Genética Forense (ISFG) se encontram a frente de cada alelo.

Os valores de parâmetros forenses, determinados para os diferentes conjuntos de marcadores na amostra populacional brasileira, se mostram muito informativos quando comparados a outras populações mundiais (Tabela 11). Os valores de Probabilidade de *Match* e Poder de Exclusão gerados pela análise do ensaio *52-plex* na população de Ribeirão Preto são equiparáveis aos resultados obtidos para um conjunto de 15 a 16 STRs. A Probabilidade de *Match* ($6,48 \times 10^{-21}$) calculada para o *52-plex* é menor quando comparada ao conjunto

PowerPlex 16 (Promega) ($5,95 \times 10^{-18}$), porém o Poder de Exclusão (0,999707) fica abaixo dos valores obtidos pelos conjuntos de STRs, até mesmo em relação ao conjunto *PowerPlex 16*, o que pode indicar eventual necessidade de análise de um maior número de SNPs em casos mais complexos de investigação de vínculos genéticos. Isso se confirma quando utilizamos para comparação os valores gerados pelo conjunto total de 139 SNPs (*52-plex* e *IISNPs*), os quais alcançam informatividade maior que as apresentadas pelos kits *GlobalFiler* e *PowerPlex Fusion*, ambos contendo 24 STRs, e maiores até que a informatividade obtida pelos 29 STRs autossômicos analisados no mesmo ensaio. Porém, deve se destacar que o Poder de Exclusão dos 92 *IISNPs*, e até mesmo dos 139 SNPs, não atinge os valores dos conjuntos de 29 STRs. Isso decorre do fato que todas as combinações de genótipos da mãe e da criança em questão (em marcadores bi-alélicos) são compatível com ao menos dois genótipos de um suposto pai, ainda que ele não seja o pai biológico.

Tabela 11. Probabilidade de *Match* e Poder de Exclusão para diferentes conjuntos de marcadores genéticos em diversas amostras populacionais de caráter global.

Conjunto de marcadores	Amostra populacional	<i>n</i>	Probabilidade de <i>Match</i> combinada	Poder de Exclusão combinado	Referência
CODIS 13	Caucasianos, afro-americanos, hispânicos e asiáticos	1036	$5,02 \times 10^{-16}$	-	(Butler et al. 2012)
CODIS 20	Caucasianos, afro-americanos, hispânicos e asiáticos	1036	$9,35 \times 10^{-24}$	-	(Butler et al. 2012)
20 STRs	Croácia	500	$3,71 \times 10^{-24}$	-	(Barbaric et al. 2017)
<i>AmpF/STR Identifier</i> (15 STRs)	Coréia do Sul	231	$2,85 \times 10^{-17}$	0,9999968	(Kim et al. 2003)
<i>AmpF/STR Identifier</i> (15 STRs)	Marrocos	209	$6,12 \times 10^{-17}$	0,9999908	(Coudray et al. 2007)
<i>PowerPlex 16 System</i> (15 STRs)	Nordeste da Espanha	204	$5,95 \times 10^{-18}$	0,999764	(Paredes et al. 2003)
16 STRs autossômicos	Maias – Guatemala	155	-	0,999991	(Cardoso et al. 2016)
<i>GlobalFiler</i> (21 STRs)	China	1006	$7,69 \times 10^{-26}$	0,999999997	(Wu et al. 2017)
<i>GlobalFiler</i> (24 STRs)	México	1008	$1,30 \times 10^{-22}$	0,99999998	(Ramos-Gonzalez et al. 2016)
<i>PowerPlex Fusion</i> (24 STRs)	México	1008	$2,35 \times 10^{-24}$	0,999999994	(Ramos-Gonzalez et al. 2016)
<i>GlobalFiler</i> (21 STRs)	Arábia Saudita	500	$1,42 \times 10^{-26}$	0,999997	(Alsafiah et al. 2017)
SNPforID 52-plex (51 SNPs)	Presente Estudo	340	$6,48 \times 10^{-21}$	0,9997	Presente Estudo
<i>IISNPs</i> (92 SNPs)	Presente Estudo	340	$4,91 \times 10^{-38}$	0,99999997	Presente Estudo
139 SNPs	Presente Estudo	340	$1,20 \times 10^{-56}$	0,99999999994	Presente Estudo
29 STRs	Presente Estudo	340	$1,19 \times 10^{-36}$	0,99999999997	Presente Estudo
13 STR do CODIS	Presente Estudo	340	$3,95 \times 10^{-16}$	0,99930	Presente Estudo
20 STR do CODIS	Presente Estudo	340	$3,99 \times 10^{-25}$	0,99997	Presente Estudo

7.4 Ancestralidade

Além de fundamentais na área de identificação humana, polimorfismos genéticos também são imprescindíveis para estimativas de ancestralidade, o que é de grande interesse na prática forense. Alguns conjuntos de marcadores específicos para esta finalidade, denominados *Ancestry Informative Markers* (AIMs) já vêm sendo desenvolvidos (Phillips et al. 2007). Adicionalmente alguns estudos utilizando marcadores do sistema *52-plex*, ainda que selecionados com o intuito de identificação, também se mostraram informativos para o estudo da estrutura populacional em populações autóctones mundiais (Sanchez et al. 2006), populações ameríndias da Venezuela (Ruiz et al. 2012) e miscigenadas do Chile (Moreno et al. 2014). Zeng et al relata em seu trabalho que a seleção destes marcadores de identificação foi realizada com banco de dados genotípicos antigos e de menor tamanho, com quantidade limitada de populações analisadas; deste modo, não se sabia no momento da seleção que alguns desses SNPs consistiam em bons preditores de ancestralidade, já que possuem frequências alélicas desiguais em alguns grupos biogeográficos (Zeng et al. 2012). Por exemplo, o SNP rs1335873 tem uma frequência alélica de 0,92/0,08 entre os africanos, de 0,28/0,71 em europeus e 0,34/0,66 em asiáticos calculados em dados do *1000Genomes* (https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=1335873).

A população de Ribeirão Preto está localizada no interior do Estado de São Paulo e é caracterizada por intensa mistura entre europeus, africanos e ameríndios. Inicialmente, a região era povoada por índios Caiapós, dispersos em aldeias por todo território. O desbravamento do sertão paulista era estimulado pela mineração. A ocupação e povoamento da região ocorreu entre o final do século XVIII e o começo do século XIX ao longo da estrada Goyases por famílias vindas de Minas Gerais, em busca de terras férteis para desenvolvimento das atividades agropecuárias. A grande maioria das terras dessa região foi ocupado por volta de 1822, de forma não legalizada e sem demarcação estipulada, o que gerava muitos conflitos entre os ocupantes (Ferreira 2014; Santos 2006). O município foi fundado em 1856, por fazendeiros que doaram parte de suas terras para a formação da cidade. Por volta de 1870, o início do desenvolvimento do cultivo do café trouxe uma grande quantidade de africanos para a região na forma de mão de obra escrava para trabalhar nas fazendas. Registros de migrações adicionais, entre os anos de 1886 e 1990, indicam um grande fluxo de imigrantes europeus, principalmente italianos, que vieram trabalhar nas

lavouras de café, devido ao aumento da necessidade de mão de obra salariada concomitante com a abolição da escravidão (1888) (Santos 2006).

Em censo realizado em 2010 pelo IBGE, tem-se para a população da cidade de Ribeirão Preto 69,77% de Brancos, 22,84% de Pardos, 6,37% de Pretos, 0,93% de Amarelos e 0,09% de Indígenas (Tabela 1). Dois estudos de ancestralidade genética utilizando marcadores STR autossômicos foram previamente realizados na região de Ribeirão Preto. Ferreira *et al.* (Ferreira *et al.* 2006), utilizando 8 STRs autossômicos em 400 indivíduos autodeclarados Brancos, determinou uma contribuição europeia, africana e ameríndia de 0,79, 0,14 e 0,07 respectivamente. Muniz *et al.* (Muniz *et al.* 2008) analisou 7 STRs em 70 indivíduos identificados como Afro-brasileiros (Pretos) e 150 indivíduos identificados como mulatos (Pardos). Para os Pardos foi obtida uma ancestralidade europeia, africana e ameríndia de 0,62, 0,26 e 0,12, respectivamente, enquanto que para os indivíduos Pretos a estimativa foi de 0,37 para contribuição europeia e 0,63 para contribuição africana, não tendo sido identificada contribuição ameríndia neste componente da população. A fim de compor os resultados de ambos os estudos e estimar uma ancestralidade total para a cidade de Ribeirão Preto, considerando a distribuição de Brancos, Pretos, Pardos, Amarelos e Indígenas descrita pelo IBGE e assumindo que Amarelos e Indígenas apresentam 100% de ancestralidade compatível com ameríndia, estimamos a média ponderada da ancestralidade genética. Deste modo, para a contribuição europeia, por exemplo, temos que a média ponderada será $(0,6977 \times 0,79) + (0,2284 \times 0,62) + (0,0637 \times 0,37) + (0,0093 \times 0,00) + (0,0009 \times 0,00) = 0,7164$. A Tabela 12 mostra os valores utilizados nos cálculos e as estimativas encontradas.

Tabela 12. Valores da contribuição ancestral de europeus, africanos e ameríndios para cada classificação étnica estabelecida pelo IBGE para a cidade de Ribeirão Preto (censo IBGE 2010), utilizados para o cálculo da estimativa total de ancestralidade baseada em STRs autossômicos.

	IBGE	Contribuição ancestral			Referências
		Europeu	Africano	Ameríndio	
Brancos	0,6977	0,79	0,14	0,07	(Ferreira <i>et al.</i> 2006)
Mulatos	0,2284	0,62	0,26	0,12	(Muniz <i>et al.</i> 2008)
Pretos	0,0637	0,37	0,63	-	(Muniz <i>et al.</i> 2008)
Amarelos	0,0093	-	-	1,00	-
Indígenas	0,0009	-	-	1,00	-
Média Ponderada		0,7164	0,1972	0,0864	

Apesar do uso de diferentes marcadores e tamanhos amostrais, as estimativas de ancestralidade encontradas no presente estudo (Tabela 5), em particular calculadas pelo

software Admix95, são equivalentes à “ancestralidade total” estimada para a cidade de Ribeirão Preto por meio de STRs autossômicos, e são consistentes com o conhecimento relacionado à história demográfica da região de Ribeirão Preto. Corroborando ainda mais os dados aqui apresentados, análise feita pelo Laboratório de Pesquisas Forenses e Genômicas utilizando o conjunto SNPforID 34-plex, composto por 34 SNPs selecionados com a finalidade de inferência de ancestralidade (Phillips et al. 2007), demonstraram estimativas de 71,4%, 19,6% e 9% de contribuição europeia, africana e ameríndia/asiática (neste caso, a parental ameríndia foi substituída por amostras populacionais do leste asiático) na composição da mesma amostra populacional utilizada no presente estudo.

Pode-se observar na Figura 3 que, mesmo com a mudança de parentais do SNPforID para as parentais do *1000Genomes*, a distribuição dos indivíduos da população amostrada não sofreu grandes alterações, porém houve um melhor agrupamento nas populações parentais quando utilizados os dados de parentais do *1000Genomes*. O melhor agrupamento das parentais também se reflete na Figura 4, onde as parentais do SNPforID apresentam certos níveis de mistura com outras parentais, o que é praticamente negligenciável nos dados do Projeto *1000Genomes*.

Por outro lado, ao analisar a inferência de ancestralidade gerada pelo *software* *STRUCTURE* apenas com o conjunto *IISNPs* (92 SNPs) nas Figura 3 e 4, percebe-se que não houve clusterização das populações parentais. Isso ocorreu devido às frequências destes marcadores entre as populações parentais serem muito próximas umas das outras, se tornando um conjunto impróprio para a realização deste tipo de análise pelo *software* *STRUCTURE*, já que o sucesso da análise depende fortemente das diferenças de frequências alélicas entre as populações parentais (Pritchard et al. 2000).

8. Conclusão

Até onde sabemos, este é o primeiro estudo que analisou todos os marcadores presentes nos conjuntos *SNPforID 52-plex*, *IISNPs* (92 SNPs), *PowerPlex Fusion* e *GlobalFiler* em uma mesma amostra populacional, o que nos permite realizar uma comparação mais fidedigna da informatividade por eles geradas.

A análise de sequenciamento de vários marcadores simultaneamente através plataforma de sequenciamento de nova geração *MiSeq Personal Sequencer* (*Illumina Inc.*) utilizando método *HaloPlex* se mostrou altamente eficiente para a geração de perfis de SNPs. Porém, ainda há falhas, principalmente relacionados aos tamanhos das *reads* geradas pelo sequenciamento, onde alelos maiores de STRs específicos podem ser encontrados com maior dificuldade, o que está relacionado à obtenção de *reads* englobando completamente as sequências de repetições e regiões flanqueadoras de interesse, necessárias para detecção realizada pelos *softwares* de análise.

O presente estudo demonstrou que a informatividade gerada por ambos os conjuntos de SNPs (*SNPforID 52-plex* e 92 SNPs) garante alta eficiência para sua utilização em casos forenses em amostra da população brasileira. Mesmo com a falta de um dos 52 SNPs nas análises, o conjunto *52-plex*, apresentou uma Probabilidade de *Match* combinada de $6,48 \times 10^{-21}$, comparável à gerada por até 16 STRs. Por outro lado, o conjunto de 92 SNPs ($PM = 4,91 \times 10^{-38}$) atingiu informatividade até mesmo maior que a gerada pelos 29 STRs autossômicos analisados neste estudo ($PM = 1,19 \times 10^{-36}$).

Por outro lado, mesmo com esta dificuldade apresentada para sequenciamento de alelos de maiores comprimentos, os STRs apresentaram como esperado alta informatividade forense. Vale destacar a grande vantagem da obtenção de perfis destes marcadores utilizando sequenciamento de nova geração, que é a possibilidade de análise nas variações presentes nas sequências de cada alelo de um mesmo tamanho, que torna os STRs ainda mais informativos, onde estudos ainda estão sendo iniciados.

As análises de ancestralidade genética demonstram que apenas o conjunto *52-plex* apresenta resultados consistentes em níveis individuais e populacional. Análise realizada com 92 SNPs mostra que estes não são bons preditores de ancestralidade devido à menor diferença

em suas frequências alélicas entre populações de diferentes ancestralidades; por outro lado, esta característica é desejada para marcadores forenses voltados para identificação, o que garante sua utilização com sucesso em diferentes populações mundiais.

Referências

- Almeida Prado Oliveira e Sousa ML, de Oliveira MA, Auler-Bittencout EA, Soares-Vieira JA, Munoz DR, e Iwamura ES. 2014. Population data of 16 autosomal STR loci of the Powerplex ESX 17 System in a Brazilian Population from the State of Sao Paulo. *Forensic Sci Int Genet* 11:e15-17.
- Alonso A, Martin P, Albarrán C, Garcia P, e Simón LF. 2005. Challenges of DNA profiling in mass disaster investigation. *Croatian Medical Journal* 46:540-548.
- Alsafiah HM, Goodwin WH, Hadi S, Alshaikhi MA, e Wepeba PP. 2017. Population genetic data for 21 autosomal STR loci for the Saudi Arabian population using the GlobalFiler((R)) PCR amplification kit. *Forensic Sci Int Genet* 31:e59-e61.
- Amigo J, Phillips C, Lareu M, e Carracedo A. 2008. The SNPforID browser: an online tool for query and display of frequency data from the SNPforID project. *Int J Legal Med* 122(5):435-440.
- Anandhakumar C, Kizaki S, Bando T, Pandian GN, and Sugiyama H. 2015. Advancing small-molecule-based chemical biology with next-generation sequencing technologies. *Chembiochem* 16(1):20-38.
- Barbaric L, Ozretic P, Horjan I, Korolija M, and Mrcic G. 2017. Forensic evaluation of the 20 STR loci in the population of Croatia. *Forensic Sci Int Genet* 28:e49-e50.
- Barbaro A, Phillips C, Fondevila M, Lareu M, and Carracedo A. 2012. Genetic variability of the SNPforID 52-plex identification SNP panel in Italian population samples. *Forensic Sci Int Genet* 6(6):e185-186.
- Bond JW. 2007. Value of DNA evidence in detecting crime. *J Forensic Sci* 52(1):128-136.
- Borsting C, Fordyce SL, Olofsson J, Mogensen HS, e Morling N. 2014. Evaluation of the Ion Torrent HID SNP 169-plex: A SNP typing assay developed for human identification by second generation sequencing. *Forensic Sci Int Genet* 12:144-154.
- Borsting C, e Morling N. 2015. Next generation sequencing and its applications in forensic genetics. *Forensic Sci Int Genet* 18:78-89.

- Børsting C, e Morling N. 2017. Use of Next-Generation Sequencing in Forensic Genetics. In: eLS. John Wiley & Sons, Ltd: ChiChester.
- Brenner C, e Morris J. 1990. Paternity index calculations in single locus hypervariable DNA probes: validation and other studies. In: Corporation P, editor. In Proceedings for the International Symposium on Human Identification 1989. Madison, WI. p 21-53.
- Budowle B, e van Daal A. 2008. Forensically relevant SNP classes. *Biotechniques* 44(5):603-608, 610.
- Butler JM. 2005. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. ACADEMIC PRESS.
- Butler JM. 2015. The future of forensic DNA analysis. *Philos Trans R Soc Lond B Biol Sci* 370(1674).
- Butler JM, Hill CR, e Coble MD. 2012. Variability of New STR Loci and Kits in US Population Groups. Disponível em: https://strbase.nist.gov/pub_pres/Profiles-in-DNA_Variability-of-New-STR-Loci.pdf (acessado 12/03/2018)
- Canturk KM, Emre R, Kinoglu K, Baspinar B, Sahin F, e Ozen M. 2014. Current status of the use of single-nucleotide polymorphisms in forensic practices. *Genet Test Mol Biomarkers* 18(7):455-460.
- Cardoso S, Sevillano R, Illescas MJ, e de Pancorbo MM. 2016. Analysis of 16 autosomal STRs and 17 Y-STRs in an indigenous Maya population from Guatemala. *Int J Legal Med* 130(2):365-366.
- Castelli Eea. 2015. HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Hum Immunol* 76(12):945-953.
- Chakraborty R. 1985. Gene identity in racial hybrids and estimation of admixture rates. In: Ahuja YR, Neel JV, editors *Genetic differentiation in human and other animal populations* Delhi: Indian Anthropological Association:171–180.
- Clayton TM, Whitaker JP, Sparkes R, e Gill P. 1998. Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Sci Int* 91:55-70.

- Coudray C, Guitard E, Keyser-Tracqui C, Melhaoui M, Cherkaoui M, Larrouy G, e Dugoujon JM. 2007. Population genetic data of 15 tetrameric short tandem repeats (STRs) in Berbers from Morocco. *Forensic Sci Int* 167(1):81-86.
- Daniel R, Santos C, Phillips C, Fondevila M, van Oorschot RA, Carracedo A, Lareu MV, e McNevin D. 2015. A SNaPshot of next generation sequencing for forensic SNP analysis. *Forensic Sci Int Genet* 14:50-60.
- del Castillo DM, Perone C, de Queiroz AR, Mourao PH, de Souza Vasconcellos L, do Nascimento MA, e Januario JN. 2009. Populational genetic data for 15 STR markers in the Brazilian population of Minas Gerais. *Leg Med (Tokyo)* 11(1):45-47.
- Dupanloup I, and Bertorelle G. 2001. Inferring admixture proportions from molecular data: Extension to any number of parental populations. *Molecular Biology and Evolution* 18(4):672-675.
- Dutta P, Bhosale S, Singh R, Gubrellay P, Patil J, Sehdev B, Bhagat S, e Bansal T. 2017. Amelogenin Gene - The Pioneer in Gender Determination from Forensic Dental Samples. *J Clin Diagn Res* 11(2):ZC56-ZC59.
- Erlich Y, Mitra PP, delaBastide M, McCombie WR, e Hannon GJ. 2008. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods* 5(8):679-682.
- Ferragut JF, Pereira R, Castro JA, Ramon C, Nogueiro I, Amorim A, e Picornell A. 2016. Genetic diversity of 38 insertion-deletion polymorphisms in Jewish populations. *Forensic Sci Int Genet* 21:1-4.
- Ferreira EB. 2014. Do caminho dos Goyases à fundação da vila de Ribeirão Preto: a questão das Sesmarias Uberaba/MG: *Boletim Jurídico*.
- Ferreira LB, Mendes-Junior CT, Wiezel CE, Luizon MR, e Simoes AL. 2006. Genomic ancestry of a sample population from the state of Sao Paulo, Brazil. *Am J Hum Biol* 18(5):702-705.
- Ferreira Palha TJB, Ribeiro Rodrigues EM, Cavalcante GC, Marrero A, de Souza IR, Seki Uehara CJ, Silveira da Motta CH, Koshikene D, da Silva DA, de Carvalho EF et al. . 2015. Population genetic analysis of insertion-deletion polymorphisms in a Brazilian population using the Investigator DIPplex kit. *Forensic Sci Int Genet* 19:10-14.

- Ganschow S, Wiegand P, e Tiemann C. 2017. toaSTR: A web-based forensic tool for the analysis of short tandem repeats in massively parallel sequencing data. *Forensic Science International: Genetics Supplement Series* 6:e119-e121.
- Gelardi C, Rockenbauer E, Dalsgaard S, Borsting C, e Morling N. 2014. Second generation sequencing of three STRs D3S1358, D12S391 and D21S11 in Danes and a new nomenclature for sequenced STR alleles. *Forensic Sci Int Genet* 12:38-41.
- Gettings KB, Aponte RA, Vallone PM, e Butler JM. 2015a. STR allele sequence variation: Current knowledge and future issues. *Forensic Sci Int Genet* 18:118-130.
- Gettings KB, Kiesler KM, e Vallone PM. 2015b. Performance of a next generation sequencing SNP assay on degraded DNA. *Forensic Sci Int Genet* 19:1-9.
- Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 11(5):759-769.
- Goodwin S, McPherson JD, e McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17(6):333-351.
- Goodwin W, Linacre A, e Hadi S. *An Introduction to Forensic Genetics*. 2 ed. John Wiley & Sons. 2011.
- Grover A, e Sharma PC. 2014. Development and use of molecular markers: past and present. *Crit Rev Biotechnol* 36(2):290-302.
- Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ et al. . 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* 48(1):22-29.
- Hagelberg E, Gray IC, e Jeffreys AJ. 1991. Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* 352:427-429.
- Hartl DL, e Clarke AG. 2007. *Principles of population genetics* A. D. Sinauer, org., Sunderland, Massachussetts: Sinauer Associates: Inc. Publishers.
- Higuchi R, von Beroldingen CH, Sensabaugh GF, e Erlich HA. 1988. DNA typing from single hairs. *Nature* 332:543-546.
- Hubisz MJ, Falush D, Stephens M, e Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour* 9(5):1322-1332.

- Hunter P. 2018. Uncharted waters: Next-generation sequencing and machine learning software allow forensic science to expand into phenotype prediction from DNA samples. *EMBO Rep* 19(3).
- Instituto Brasileiro de Geografia e Estatística. IBGE. 2010. Censo Demográfico, Rio de Janeiro. Disponível em: <http://censo2010.ibge.gov.br/>. (acessado em 09/07/2018).
- Illumina. 2013. Preparing Libraries for Sequencing on the MiSeq. California, USA. p 14.
- International SNP Map Working Group. 2011. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928-933.
- Jakovski Z, Ajanovska RJ, Stankov A, Poposka V, Bitoljanu N, e Belakaposka V. 2017. The power of forensic DNA data bases in solving crime cases. *Forensic Science International: Genetics Supplement Series* 6:e275-e276.
- Jeffreys AJ, Allen JM, Hagelberg E, e Sonnberg A. 1992. Identification of the skeletal remains of Josef Mengele by DNA analysis. *Forensic Sci Int* 56:65-76.
- Jeffreys AJ, Wilson V, e Thein SL. 1985. Individual-specific 'fingerprints' of human DNA. *Nature* 316:76-79.
- Jones DA. 1972. Blood samples: Probability of discrimination. *Forensic Sci Soc* 12:355-359.
- Kayser M, e Schneider PM. 2009. DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. *Forensic Sci Int Genet* 3(3):154-161.
- Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SL, Karoma NJ, Kungulilo S, Kim JJ, Lu RB, Odunsi A et al. . 2006. Developing a SNP panel for forensic identification of individuals. *Forensic Sci Int* 164(1):20-32.
- Kim YL, Hwang JY, Kim YJ, Lee S, Chung NG, Goh HG, Kim CC, e Kim DW. 2003. Allele frequencies of 15 STR loci using AmpF/STR identifier kit in a Korean population. *Forensic Sci Int* 136(1-3):92-95.
- Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, Paul P, Albrecht V, Andreas JM, Baier DM et al. . 2014. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics* 15(63).
- Lau E, Mohandas T, Shapiro L, Siavkin H, e Snead M. 1989. Human and mouse amelogenin gene loci are on the sex chromosomes. *Genomics* 4:162-168.

- Lee SB, McCord B, e Buel E. 2014. Advances in forensic DNA quantification: a review. *Electrophoresis* 35(21-22):3044-3052.
- Li H, e Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
- Lima TH, Buttura RV, Donadi EA, Veiga-Castelli LC, Mendes-Junior CT, e Castelli EC. 2016. HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample. *Hum Immunol* 77(10):841-853.
- Lotte Hansen L, Eskerod Madsen B, Pedersen K, e Wiuf C. 2007. Conflicting results in SNP genotype assessment. *Biotechniques* 43(6):756-762.
- Mannucci A, Sullivan KM, Ivanov PI, e Gill P. 1994. Forensic application of a rapid and quantitative DNA sex test by amplification of the X-Y homologous gene amelogenin. *International Journal of Legal Medicine* 106:190-193.
- Martim M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17(1):10-11.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. . 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297-1303.
- Mendes-Junior, CT. Técnicas aplicadas à análise forense de DNA. In: Bruni AT, Velho JA, e Oliveira, MF (Orgs). *Fundamentos de Química Forense: Uma análise prática da química que soluciona crimes*. Campinas: Millenium, 2012. 191-207 p.
- Metzker ML. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11(1):31-46.
- Miller Sa, Dykes DD, e Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research* 16(3):1215.
- Ministério da Justiça e Cidadania. III RELATÓRIO DA REDE INTEGRADA DE BANCOS DE PERFIS GENÉTICOS (RIBPG). 2015
- Ministério da Justiça e Cidadania. VII RELATÓRIO DA REDE INTEGRADA DE BANCOS DE PERFIS GENÉTICOS (RIBPG). 2017

- Moreno F, Freire-Aradas A, Phillips C, Fondevila M, Carracedo A, e Lareu MV. 2014. SNP variation with latitude: Analysis of the SNPforID 52-plex markers in north, mid-region and south Chilean populations. *Forensic Sci Int Genet* 10:12-16.
- Muniz YC, Ferreira LB, Mendes-Junior CT, Wiezel CE, e Simoes AL. 2008. Genomic ancestry in urban Afro-Brazilians. *Ann Hum Biol* 35(1):104-111.
- Musgrave-Brown E, Ballard D, Balogh K, Bender K, Berger B, Bogus M, Borsting C, Brion M, Fondevila M, Harrison C et al. . 2007. Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet* 1(2):186-190.
- Nunotani M, Shiozaki T, Sato N, Kamei S, Takatsu K, Hayashi T, Ota M, e Asamura H. 2015. Analysis of 30 insertion-deletion polymorphisms in the Japanese population using the Investigator DIPplex(R) kit. *Leg Med (Tokyo)* 17(6):467-470.
- Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, e Kidd KK. 2010. SNPs for a universal individual identification panel. *Hum Genet* 127(3):315-324.
- Paredes M, Crespillo M, Luque JA, e Valverde JL. 2003. STR frequencies for the PowerPlex (R) 16 system kit in a population from Northeast Spain. *Forensic Sci Int* 135(1):75-78.
- Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmao L, Hares DR, Irwin JA, King JL et al. . 2016. Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Sci Int Genet* 22:54-63.
- Pepinski W, Abreu-Glowacka M, Koralewska-Kordel M, Michalak E, Kordel K, Niemcunowicz-Janica A, Szeremeta M, e Konarzewska M. 2013. Population genetics of 30 INDELs in populations of Poland and Taiwan. *Mol Biol Rep* 40(7):4333-4338.
- Phillips C, Fondevila M, Garcia-Magarinos M, Rodriguez A, Salas A, Carracedo A, e Lareu MV. 2008. Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers. *Forensic Sci Int Genet* 2(3):198-204.
- Phillips C, Salas A, Sanchez JJ, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV et al. . 2007. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet* 1(3-4):273-280.

- Poulsen L, Borsting C, Tomas C, Gonzalez-Andrade F, Lopez-Pulles R, Gonzalez-Solorzano J, e Morling N. 2011. Typing of Amerindian Kichwas and Mestizos from Ecuador with the SNPforID multiplex. *Forensic Sci Int Genet* 5(4):e105-107.
- Pritchard JK, Stephens M, e Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2):945-959.
- Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, e Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13:341.
- Ramos-Gonzalez B, Aguilar-Velazquez JA, Chavez-Briones Mde L, Delgado-Chavarria JR, Alfaro-Lopez E, e Rangel-Villalobos H. 2016. Population data of 24 STRs in Mexican-Mestizo population from Monterrey, Nuevo Leon (Northeast, Mexico) based on Powerplex((R)) Fusion and GlobalFiler((R)) kits. *Forensic Sci Int Genet* 21:e15-17.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, e Mesirov JP. 2011. Integrative Genomics Viewer. *Nature Biotechnology* 29:24-26.
- Rockenbauer E, Hansen S, Mikkelsen M, Borsting C, e Morling N. 2014. Characterization of mutations and sequence variants in the D21S11 locus by next generation sequencing. *Forensic Sci Int Genet* 8(1):68-72.
- Rousset F. 2008. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour* 8(1):103-106.
- Ruiz Y, Chiurillo MA, Borjas L, Phillips C, Lareu MV, e Carracedo A. 2012. Analysis of the SNPforID 52-plex markers in four Native American populations from Venezuela. *Forensic Sci Int Genet* 6(5):e142-145.
- Sanchez JJ, Phillips C, Borsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, Salas A, Syndercombe-Court D et al. . 2006. A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27(9):1713-1724.
- Sanger F, Nicklen S, e Coulson R. 1997. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74(12):5463-5467.

- Santos C, Phillips C, Fondevila M, Porras-Hurtado L, Carracedo A, Souto L, e Lareu MV. 2011. A study of East Timor variability using the SNPforID 52-plex SNP panel. *Forensic Sci Int Genet* 5(1):e25-26.
- Santos FR. 1998. Reliability of DNA-based sex tests. *Nature* 18:103.
- Santos JR. 2006. Imigração e Ascensão Social em Ribeirão Preto entre o final do século XIX e meados do XX.
- SNPforID. <http://spsmart.cesga.es/snpforid.php?dataSet=snpforid52>.
- Srivastava M, Tripathi S, Astekar M, Singal D, Srivastava A, e Vashisth P. 2017. Sex determination from mesiodens of Indian children by amelogenin gene. *J Forensic Dent Sci* 9(3):125-129.
- Steinbock LJ, e Radenovic A. 2015. The emergence of nanopores in next-generation sequencing. *Nanotechnology* 26(7):074003.
- STRBase. <http://www.cstl.nist.gov/strbase/index.htm>.
- Thorvaldsdóttir H, Robinson JT, e Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14(2):178-192.
- van Dijk EL, Auger H, Jaszczyszyn Y, e Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet* 30(9):418-426.
- Van Neste C, Van Nieuwerburgh F, Van Hoofstat D, e Deforce D. 2012. Forensic STR analysis using massive parallel sequencing. *Forensic Sci Int Genet* 6(6):810-818.
- van Oorschot RA, Ballantyne KN, e Mitchell RJ. 2010. Forensic trace DNA: a review. *Investig Genet* 1(1):14.
- Vilsen SB, Tvedebrink T, Eriksen PS, Bosting C, Hussing C, Mogensen HS, e Morling N. 2018. Stutter analysis of complex STR MPS data. *Forensic Sci Int Genet* 35:107-112.
- Warshauer DH, King JL, e Budowle B. 2015. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. *Forensic Sci Int Genet* 14:182-186.
- Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, King JL, e Budowle B. 2013. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet* 7(4):409-417.

- Wei Y-L, Qin C-J, Liu H-B, Jia J, Hu L, e Li C-X. 2014. Validation of 58 autosomal individual identification SNPs in three Chinese populations. *Croatian Medical Journal* 55(1):10-13.
- Wendt FR, Zeng X, Churchill JD, King JL, e Budowle B. 2016. Analysis of Short Tandem Repeat and Single Nucleotide Polymorphism Loci From Single-Source Samples Using a Custom HaloPlex Target Enrichment System Panel. *Am J Forensic Med Pathol* 37(2):99-107.
- Woerner AE, King JL, e Budowle B. 2017. Fast STR allele identification with STRait Razor 3.0. *Forensic Sci Int Genet* 30:18-23.
- Wu L, Pei B, Ran P, e Song X. 2017. Population genetic analysis of Xiamen Han population on 21 short tandem repeat loci. *Leg Med (Tokyo)* 26:41-44.
- Yang X, Liu D, Liu F, Wu J, Zou J, Xiao X, Zhao F, e Zhu B. 2013. HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics* 14:33.
- Yang Y, Xie B, e Yan J. 2014. Application of next-generation sequencing technology in forensic science. *Genomics Proteomics Bioinformatics* 12(5):190-197.
- Zeng Z, Wang L, Feng Q, Zhang L, Lee L, Wang L, Yue Y, Fang Y, Yang W, Qiu H et al. . 2012. Evaluation of 96 SNPs in 14 populations for worldwide individual identification. *J Forensic Sci* 57(4):1031-1035.
- Zhang S, Bian Y, Chen A, Zheng H, Gao Y, Hou Y, e Li C. 2017. Massively parallel sequencing of 231 autossomal SNPs with a custom panel: a SNP typing assay developed for human identification with Ion Torrent PGM. *Forensic Sciences Research*.
- Zhang S, Niu Y, Bian Y, Dong R, Liu X, Bao Y, Jin C, Zheng H, e Li C. 2018. Sequence investigation of 34 forensic autosomal STRs with massively parallel sequencing. *Sci Rep* 8(1):6810.

Apêndice A – Protocolo de processamento computacional para genotipagem de STRs

Laboratório de Pesquisas Forenses e Genômicas
Departamento de Química – FFCLRP/USP

Protocolo de processamento computacional de dados de sequenciamento de nova geração para genotipagem de STRs

Versão 1 – Julho de 2018

ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO PARA GENOTIPAGEM DE STRs	97
1.1 Programas	97
1.1.1 <i>toaSTR</i>	97
1.1.2 <i>STRait Razor</i>	100
1.2 Regras para confirmação de genótipos	102
1.3 Referências	104

ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO PARA GENOTIPAGEM DE STRs

Duas ferramentas de identificação de alelos de STR em sequenciamento de nova geração, *toaSTR* (Ganschow et al. 2017) e *Strait Razor* (King et al. 2017; Warshauer et al. 2015; Warshauer et al. 2013; Woerner et al. 2017), podem ser empregadas para a obtenção de genótipos. Ambas as ferramentas apresentam métodos que se baseiam na identificação das sequências flanqueadoras da região de repetição para a detecção dos alelos de cada STR.

Para análise de STRs, os arquivos utilizados são os gerados após a retirada (trimagem através do programa *cutadapt*), das sequências indesejadas (inseridas durante o sequenciamento), presentes nos dados gerados pelo sequenciamento como descrito detalhadamente no Protocolo de processamento computacional de dados de sequenciamento de nova geração - versão 2 (Anexo D), do Laboratório de Pesquisas Forenses e Genômicas.

1.1 Programas

1.1.1 *toaSTR*

toaSTR é uma ferramenta de análise de STR disponível na *web* de livre acesso, apenas para fins de pesquisa até o momento. Para utilizá-la, é necessário criar um login e senha de acesso. Os dados a serem analisados independem de kits e plataforma de sequenciamento utilizada. Todo suporte e detalhes de funcionamento se encontram no manual disponibilizado no site <https://toastr.de>.

1.1.1.1 Método de análise

Simplificadamente, o *toaSTR* emprega o pacote BBTools para pré-processamento de dados. O algoritmo valida o arquivo de entrada adicionado e realiza uma pesquisa difusa rápida em todas as *reads* geradas no sequenciamento. A *read* que apresentar elementos de reconhecimentos internos (RE, do inglês *Recognition Elements*) é selecionada para um sistema de STR. Esses REs nada mais são do que as regiões flaqueadoras, a jusante e montante da região de repetições, específicas para cada *locus* de STR. O padrão do algoritmo aceita até duas incompatibilidades (*mismatches*) em cada RE. As *reads* que não contiverem a região de repetição dos STRs do sistema, são descartadas.

Na análise de um STR de uma dada amostra, as *reads* contendo regiões de repetição idênticas são agrupadas, suas respectivas frequências estabelecidas (cobertura) e, de acordo com parâmetros estabelecidos pelo usuário (vistos a seguir), seguem para a etapa de alinhamento com a estrutura de repetição. É ainda determinado o comprimento do fragmento para equivalência com o procedimento de eletroforese capilar (EC). *toastr* conta também com um recurso de modelagem de *stutter* integrado que classifica as observações obtidas em três categorias: alelos, *stutter* ou artefato.

1.1.1.2 Etapas para análises

A) Configuração do painel de marcadores

Primeiro passo é definir quais marcadores serão analisados. Na guia Painel (*Panels*), escolhe-se os marcadores dentre os 24 STR autossômicos (D1S1656, D2S1338, D2S441, TPOX, D3S1358, FGA, CSF1PO, D5S818, D6S1043, SE33, D7S820, D8S1179, D10S1248, TH01, D12S391, VWA, D13S317, PENTA E, D16S539, D18S51, D19S433, D21S11, PENTA D, D22S1045), um STR do cromossomo Y (DYS391), e ainda a inserção/deleção do gene da Amelogenina, voltada para tipagem do sexo. Neste mesmo painel adicionamos valores de corte para determinação de *stutter* (*Stutter Threshold*), para cada *locus* individualmente.

Para este estudo, todos os 26 marcadores presentes no *software* foram utilizados na análise.

B) Iniciando nova análise:

Para se iniciar uma nova análise, deve-se seguir os seguintes passos:

1º Passo – Nomear amostra.

2º Passo – Escolher tipo de material analisado (sangue, saliva, pele, etc...).

3º Passo – Determinar se amostra é de possível mistura ou amostra sabidamente individual.

4º Passo – Escolher o painel de marcador (previamente configurado) a ser utilizado.

5º Passo – Inserir valor de *Analytical Threshold* (AT).

6º Passo – Inserir valor de *Calling Threshold* (CT).

7º Passo – Fazer *upload* do arquivo FASTQ das amostras, contendo um tamanho máximo de 500 megabytes.

As análises são realizadas para uma amostra de cada vez.

Tabela A1. Definição e valores dos parâmetros utilizados no presente estudo para análises no toaSTR.

Parâmetro	Abreviação	Valores	Descrição
<i>Mismatches</i>	Mm	2 (<i>default</i>)	Mutações, inserções, substituições presentes nas regiões flangeadoras
<i>Stutter Threshold</i>	ST	15%	Porcentagem para que determinado pico seja chamado de <i>stutter</i>
<i>Calling Threshold</i>	CT	2	Mínimo de cobertura para chamar alelo de verdadeiro
<i>Analytical Threshold</i>	AT	1	Cobertura total mínima de <i>reads</i>

C) Visualização e obtenção dos resultados

Os resultados de cada análise ficam salvos e podem ser interpretadas a qualquer tempo. No resumo dos resultados da análise de todas as amostras, encontramos informações de cada amostra bem como as configurações e parâmetros utilizados.

A escolha de uma amostra possibilita a análise individual, onde a tabela genotípica é exibida, fornecendo os possíveis genótipos (concordantes com a nomenclatura da CE, de acordo com seu comprimento) de cada marcador (STR) previamente selecionado no painel. Adicionalmente, ao selecionarmos um marcador, temos um resumo detalhado dos resultados, com a ilustração tabular e gráfica dinâmica. A tabela a ser exibida, mostra a sequência de cada alelo detectado, bem como seu *status* (alelo, *stutter* ou artefato) seguido da cobertura detectada. Estes dados são ilustrados na forma gráfica, facilitando sua análise e interpretação. A visualização desses dados é possível apenas na plataforma *online*.

O *toaSTR* permite ainda exportar estes dados em um arquivo no formato .PDF, contendo informações condensadas da análise e todas as informações referentes a cada marcador (nome do *locus*, o nome do alelo de acordo com CE, o cromossomo, a versão do

genoma humano de referência, as coordenadas da região repetitiva do alelo de referência e a descrição dos *motifs* STR). Por fim, um arquivo voltado para cálculos bioestatísticos pode ser gerado, no formato CSV, contendo apenas os alelos de cada *locus*.

Após todas as amostras analisadas, os genótipos gerados no arquivo .csv foi conferido manualmente a fim de sua confirmação, e todos os resultados foram reunidos em um único arquivo para dar procedimento nos próximos passos.

1.1.2 STRait Razor

Diferentemente da ferramenta *toaSTR*, o *software STRait Razor* deve ser instalado em computador local para sua utilização e encontra-se disponível para download em <https://www.unthsc.edu/graduate-school-of-biomedical-sciences/molecular-and-medical-genetics/laboratory-faculty-and-staff/strait-razor/>. Desde seu desenvolvimento, o *software* vem passando por constantes melhorias tanto no quesito de velocidade de análise, quanto nas ferramentas de leituras e interpretação dos resultados (King et al. 2017; Warshauer et al. 2015; Warshauer et al. 2013). Sua versão v3 expandiu suas funcionalidades para *Windows*, podendo, a partir de então, ser utilizado em qualquer sistema operacional (*Windows, Mac ou Linux*)(Woerner et al. 2017).

1.1.2.1 Método de análise

Ambos os programas utilizados apresentam os mesmos princípios para detecção dos marcadores nos dados de sequenciamento de nova geração. As *reads* contendo ambas as regiões flangeadoras das sequencias de repetições são selecionadas, permitido a existência de *mismatches* (*default* 2), ou seja, bases presentes nas regiões flangeadoras que podem ser diferentes das presente na sequência referência. Toda região do marcador, incluindo as sequências flangeadoras e a região de repetição, é selecionada. As regiões de repetição são então filtradas com base na presença de uma pequena porção das unidades de repetição. Em seguida, o número de bases em cada região de repetição filtrada é determinado. A determinação do alelo é realizada pela comparação de regiões de repetição encontradas (realizada a contagem do número de bases presentes) com as de comprimentos conhecidos (Warshauer et al. 2013).

1.1.2.2 Etapas para análises

A) Dinâmica de análise

O pacote completo de arquivos e manuais necessários para utilização da versão para o *Windows* do *STRait Razor v3*, incluindo planilhas de análise dos resultados, estão disponibilizado em <https://github.com/Ahhgust/STRaitRazor>.

Quando realizado o download do pacote *STRait Razor*, ele já nos fornece um arquivo *batch*. Arquivos *batch* são simplesmente arquivos de textos, onde cada linha contém comandos definidos em um *script*. Assim, o arquivo *batch script*, contém todas as linhas de comandos necessárias para o funcionamento do *software*. Para inserir os marcadores a serem analisados, deve-se alterar a linha 5 (*SET config*) deste arquivo, modificando o caminho pré-existente para o de um arquivo de configuração desejado, contendo as informações todos os marcadores que deverão ser analisados pelo *software*. O mínimo para que um *locus* seja levado em consideração envolve cobertura de no mínimo 2 *reads* (AT=2) e cobertura de no mínimo dois para se aceitar a existência de um alelo (CT=2). Esta planilha ainda permite verificar as sequências de cada alelo, inserir novas nomenclaturas, dentre outras ferramentas. Para inicializar a análise, basta executar o arquivo *batch script*, e a análise de todas as amostras presentes na mesma pasta em que ele se encontra será realizada.

O *software STRait Razor* nos permite inserir marcadores para análise além daqueles já presentes por padrão no programa. Deste modo, para que se pudesse realizar a análise de todos os STRs presentes no sequenciamento realizado pelo LPFG, houve a necessidade da adição de 6 STRs que inicialmente não eram analisados pelo *software*, no banco de dados da planilha. Todos os passos para inclusão de qualquer STRs são descritos no *STRait Razor Analysis Manual*, disponível em <https://www.unthsc.edu/graduate-school-of-biomedical-sciences/wp-content/uploads/sites/13/STRait-Razor-Analysis-Manual.pdf>. Para realizar a adição, deve-se levar em consideração regiões flanqueadoras obtidas pela observação no *software* IGV. Deve-se buscar obter um mínimo de 12 nucleotídeos em cada flanqueadora, visando uma maior especificidade. Além disso, informações do número de nucleotídeos presente em cada unidade de repetição (Período), bem como um pedaço contendo unidade de repetições (*Motif*) devem ser fornecidas. A Tabela 2 mostra o formato do arquivo de entrada inserindo os STRs, com suas respectivas informações necessárias.

Tabela A2. Informações inseridas para captura de outros STRs analisados pelo LPFG pelo programa *STrait Razor v3*.

Marcador	Tipo	Regiões flanqueadoras	Motif	Período	Offset*
F13A01	Autossômico	[AATTTTTTTCTTTTACT]... [ATCTTTCTATCTTTTCAGATGAA]	TTTCTTTCTTTC	4	1
F13B	Autossômico	[AAGACTCTGTCTCT]... [TTCCAAGGGTCA]	AAATAAATAAAT	4	14
FES/FPS	Autossômico	[AAGGATAGCAGTTC]... [GAGATGTAGTCTCATT]	ATTTATTTATTT	4	1
LPL	Autossômico	[AACATTCCAAGAATTA]... [TTTTTGAGACACAGT]	TTTATTTATTTA	4	1
PentaC	Autossômico	[CTAGGGACTTCTT]... [AGTCCCTCTTTTC]	AAAACAAAAC	5	6
SE33	Autossômico	[ACCGCTATAGTAA]... [TCTTCCTCCTTTCTCTCTCT]	TTTCTTTC	4	1

*número de nucleotídeos presentes entre as regiões flanqueadoras e as unidades de repetições, que devem ser desconsiderados para determinação dos alelos.

B) Visualização dos resultados

O *software Strait Razor* gera um arquivo de texto no formato TXT como *output*. Este arquivo contém o nome de cada *locus* e os alelos encontrados, com suas respectivas sequências e coberturas. Para visualização e organização destes dados, os desenvolvedores do *software* disponibilizam uma planilha no *software Microsoft Excel* (chamada *Strait Razor Analysis v3*), onde pode-se analisar todos os arquivos de amostras simultaneamente e gerar uma tabela organizada com genótipos e cobertura de cada alelo encontrado.

1.2 Regras para confirmação de genótipos

Devido à baixa cobertura de alguns *loci*, pode-se gerar genótipos que não sejam verdadeiros. Para corrigir essa denominação falso-positiva, utiliza-se para a determinação dos genótipos de ambas as ferramentas, as seguintes regras (não levando em consideração mutações de pontos presentes dentro das regiões de repetição):

- A homozigose é aceita em situações caracterizadas por genótipos com um único alelo quando é alcançada uma cobertura mínima de oito *reads*. Este procedimento assegura ($p > 0,99$) que um genótipo homozigoto é chamado por causa da falta de variância nessa posição e não devido a não amostragem do segundo alelo.

- A homozigose também é aceita em situações caracterizadas por genótipos com dois alelos adjacentes (ex: alelo 16 e alelo 15), em que o alelo menor é sub representado (proporção de *reads* 15%), sendo considerado como homozigoto para o alelo de maior tamanho. Este procedimento minimiza a influência do *stutter* que caracteriza a amplificação por PCR de alelos de STRs.
- Homozigose também é aceita em situações caracterizadas por genótipos com dois alelos, em que um alelo foi extremamente sub representado (proporção de *reads* abaixo de 1%), sendo considerada como homozigotos para o alelo mais representado.
- Heterozigose que envolve dois alelos adjacentes com o alelo menor sendo sub representado só é aceita quando uma cobertura mínima de oito *reads* for alcançada e a proporção de *reads* para o alelo menor for superior a 15%. Se esta condição não for satisfeita, é introduzido um alelo interrogado no lugar do alelo inferior sub representado. Este procedimento minimiza a probabilidade de amplificação de *stutter*, resultando em um falso heterozigoto quando o indivíduo poderia ser (a) homozigoto para o alelo maior observado ou (b) heterozigoto incluindo o alelo maior e um alelo diferente não amostrado pelo acaso.
- Heterozigose que envolva (a) alelos não adjacentes ou (b) alelos adjacentes com o alelo menor sendo igualmente ou sobre representado, foi aceito somente quando havia cobertura mínima de quatro *reads* e a proporção de *reads* para o alelo menos representado era superior a 25%. Se estas condições não forem satisfeitas, um alelo interrogado é introduzido no lugar do alelo sub representado (ou ambos, no caso da ocorrência de um único alelo numa situação de cobertura igual a dois). Este procedimento assegura que apenas os genótipos heterozigotos de alta qualidade são chamados e é particularmente útil em situações caracterizadas por baixa cobertura, na qual algumas *reads* podem indicar a existência de um alelo alternativo; tais leituras podem incluir erros de sequenciamento (variante falso positiva) ou podem representar um genótipo verdadeiro associado a desequilíbrio de heterozigotos (variante verdadeira positiva).

1.3 Referências

- Ganschow S, Wiegand P, e Tiemann C. 2017. toaSTR: A web-based forensic tool for the analysis of short tandem repeats in massively parallel sequencing data. *Forensic Sci Int: Genetics Supplement Series* 6:e119-e121.
- King JL, Wendt FR, Sun J, e Budowle B. 2017. STRait Razor v2s: Advancing sequence-based STR allele reporting and beyond to other marker systems. *Forensic Sci Int. Genet* 29:21-28.
- Warshauer DH, King JL, e Budowle B. 2015. STRait Razor v2.0: the improved STR Allele Identification Tool--Razor. *Forensic Sci Int Genet* 14:182-186.
- Warshauer DH, Lin D, Hari K, Jain R, Davis C, Larue B, King JL, e Budowle B. 2013. STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet* 7(4):409-417.
- Woerner AE, King JL, e Budowle B. 2017. Fast STR allele identification with STRait Razor 3.0. *Forensic Sci Int Genet* 30:18-23.

Apêndice B – Populações do SNPforID e Projeto 1000Genomes utilizadas na inferência de ancestralidade

Tabela B1. Populações parentais utilizadas para cálculos de ancestralidade, obtidas à partir dos dados do SNPforID browser (SNPforID).

Nome	Abreviação	<i>n</i>	Classificação
Alemanha	GER	49	Europeia
Itália - Reggio Calabria	Ita-RC	33	Europeia
Itália - Veneto	Ita-V	30	Europeia
Itália - Sardinian	Ita-S	28	Europeia
Itália - Tuscan	Ita-T	8	Europeia
Itália - Bergamo	Ita-B	13	Europeia
Espanha	Spain	151	Europeia
Portugal	Port	193	Europeia
Angola	Ango	80	Africano
C. African Republic - Bianca Pigmy	CABP	23	Africano
D. R. of Congo - Mbuti Pygmy	CoMP	13	Africano
Kenya - Bantu N.E.	KenB	11	Africano
Mozambique	Moza	153	Africano
Namibia - San	Nsan	6	Africano
Nigeria - Youba	NigY	22	Africano
Nigeria - resident UK	NigUK	10	Africano
Senegal - Mandenka	SenM	22	Africano
Somalia	Soma	104	Africano
South Africa - Bantu S. E.	SABa	8	Africano
Uganda	Ugan	60	Africano
Brasil - Karitiana	BR-K	14	Ameríndio
Brazil - Surui	BR-S	8	Ameríndio
Colombia - Awa	Col-A	36	Ameríndio
Colombia - Coyaima	Col-C	62	Ameríndio
Colombia - Embera	Col-E	18	Ameríndio
Colombia - Pijao	Col-P	35	Ameríndio
Ecuador - Kichwa	Equa-K	77	Ameríndio
Total de indivíduos		1267	

Disponível em :<http://spsmart.cesga.es/snpforid.php?dataSet=snpforid52>

Tabela B2. Populações parentais utilizadas para cálculos de ancestralidade, obtidas a partir dos dados *1000Genomes*.

Nome	Abreviação	<i>n</i>	Classificação
Finlândia	FIN	99	Europeia
Britânicos – Inglaterra e Escócia	GBR	91	Europeia
Itália – Toscana	TSI	107	Europeia
Ibéricos - Espanha	IBS	107	Europeia
Nigéria – Yoruba de Ibadan	YRI	108	Africano
Quênia – Quos Luhya de Webuye	LWK	99	Africano
Gambia – Divisão Leste	GWD	113	Africano
Nigéria - Esan	ESN	99	Africano
Serra Leoa - Mende	MSL	85	Africano
China – Han de Pequim	CHB	103	Asiático
Japão - Tóquio	JPT	104	Asiático
China – Dai de Xishuangbanna	CDX	93	Asiático
China – Han do Sul da China	CHS	105	Asiático
Vietnã – Kinh in Ho Chi Minh	KHV	99	Asiático
Total de indivíduos		1412	

Disponível em: <http://www.internationalgenome.org/>

Apêndice C – Tabelas de comparações de frequências alélicas

As tabelas abaixo apresentam os valores de frequências alélicas de 28 dos 29 STRs autossômicos analisados no presente estudo, comparando resultados obtidos em diferentes estudos realizados no Brasil, com ênfase em populações da região Sudeste. Os valores 0,000 indicam que o alelo em questão foi observado com frequência abaixo de 0,1% (0,001) na respectiva população.

Tabela C1. Frequências alélicas no *locus* CSF1PO.

Alelos	BR1	BR2	RJ	MS	MG	SP1	RP
5	0,000						
6	0,001	0,000			0,001		
7	0,013	0,010	0,043	0,029	0,019	0,001	0,016
8	0,020	0,015	0,021	0,010	0,023	0,025	0,021
9	0,033	0,014	0,030	0,024	0,024	0,024	0,038
10	0,262	0,262	0,278	0,250	0,278	0,022	0,276
10.3		0,000					
11	0,294	0,298	0,218	0,327	0,287	0,269	0,304
12	0,302	0,332	0,368	0,288	0,301	0,275	0,282
13	0,064	0,059	0,043	0,072	0,059	0,310	0,056
14	0,010	0,008			0,007	0,067	0,006
15	0,001	0,002			0,001	0,006	
16	0,000						
22	0,000						
23	0,000						

BR1:Brasil (Aguiar et al. 2012); BR2:Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP1: São Paulo (Fridman et al. 2008); RP: Ribeirão Preto (Presente estudo)

Tabela C2. Frequências alélicas no *locus* D1S1656.

Alelos	BR2	SP2	RP
8		0,000	
10	0,005	0,003	0,006
11	0,067	0,055	0,055
12	0,096	0,099	0,101
13	0,083	0,087	0,073
14	0,126	0,145	0,155
14.3	0,006	0,006	0,009
15	0,145	0,154	0,165
15.3	0,044	0,045	0,043
16	0,126	0,141	0,116
16.2		0,000	
16.3	0,064	0,048	0,052
17	0,037	0,047	0,038
17.1	0,000		

Continua

<i>Conclusão</i>			
Alelos	BR2	SP2	RP
17.3	0,140	0,111	0,120
18	0,004	0,003	0,006
18.3	0,044	0,046	0,052
19	0,001	0,000	
19.2			0,003
19.3	0,009	0,008	0,003
20.3	0,002	0,000	0,003

BR2: Brasil (Moyses et al. 2017); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo)

Tabela C3. Frequências alélicas no *locus* D2S1338.

Alelos	BR2	MG	SP2	RP
12			0,000	
13			0,001	
14	0,002		0,001	
15	0,000	0,001		0,003
16	0,044	0,052	0,042	0,074
17	0,217	0,204	0,224	0,213
18	0,077	0,079	0,086	0,091
19	0,132	0,118	0,115	0,111
20	0,135	0,123	0,127	0,139
21	0,051	0,063	0,049	0,042
22	0,074	0,080	0,054	0,070
23	0,117	0,100	0,110	0,126
24	0,073	0,089	0,092	0,069
24.2	0,000			
25	0,059	0,071	0,079	0,050
26	0,017	0,017	0,014	0,008
27	0,001	0,002	0,004	0,003
28		0,000	0,002	
29		0,000		

BR2:Brasil (Moyses et al. 2017); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C4. Frequências alélicas no *locus* D2S441.

<i>Continua</i>			
Alelos	BR2	SP2	RP
8	0,000		0,003
9	0,002	0,002	0,003
9.1		0,001	
10	0,263	0,221	0,223
11	0,306	0,308	0,314
11.3	0,051	0,056	0,061
12	0,057	0,071	0,060
12.3	0,005	0,005	0,003
13	0,026	0,028	0,024
13.1			0,001
13.3	0,000	0,000	0,001
14	0,253	0,263	0,269
14.3		0,000	

Conclusão			
Alelos	BR2	SP2	RP
15	0,033	0,043	0,036
16	0,003	0,002	
17			0,001

BR2: Brasil (Moyses et al. 2017); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C5. Frequências alélicas no *locus* D3S1358.

Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
6	0,000						
7	0,000						
8	0,000						
9	0,000	0,001			0,000		
10	0,000				0,000	0,000	
11	0,001	0,000			0,001		
12	0,002	0,004		0,010	0,004	0,002	
13	0,004	0,004	0,014	0,005	0,003	0,002	0,003
14	0,093	0,078	0,110	0,064	0,094	0,091	0,106
15	0,291	0,314	0,310	0,294	0,281	0,279	0,276
15.2	0,000	0,000			0,000	0,000	
16	0,270	0,283	0,329	0,255	0,274	0,257	0,255
16.2	0,000				0,000		
16.3	0,000						
16.5	0,000						
17	0,203	0,193	0,162	0,186	0,210	0,219	0,211
17.3	0,000						
18	0,124	0,112	0,076	0,181	0,122	0,138	0,139
19	0,010	0,008		0,005	0,009	0,010	0,009
20	0,001	0,002			0,001	0,000	0,001
21	0,000	0,000					

BR1:Brasil (Aguiar et al. 2012); BR2:Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C6. Frequências alélicas no *locus* D5S818.

Continua							
Alelos	BR1	BR2	RJ	MS	MG	SP1	RP
5	0,000						
6	0,000				0,000	0,001	
7	0,017	0,034	0,017	0,024	0,011	0,013	0,012
8	0,018	0,014	0,030	0,024	0,025	0,021	0,024
9	0,038	0,043	0,026	0,024	0,034	0,053	0,034
10	0,062	0,055	0,073	0,067	0,061	0,057	0,058
11	0,329	0,332	0,308	0,365	0,311	0,307	0,294
12	0,346	0,339	0,338	0,361	0,352	0,336	0,374
13	0,174	0,169	0,188	0,115	0,192	0,203	0,192
14	0,013	0,013	0,021	0,019	0,012	0,010	0,012
15	0,002	0,001			0,002		
16	0,000				0,000		
17	0,000	0,000					
18	0,000						

Alelos	BR1	BR2	RJ	MS	MG	SP1	RP
19	0,000						

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP1: São Paulo (Fridman et al. 2008); RP: Ribeirão Preto (Presente estudo).

Tabela C7. Frequências alélicas no *locus* D7S820.

Alelos	BR1	BR2	RJ	MS	MG	SP1	RP
5	0,000						
6	0,000		0,004		0,000		
6.2	0,000						
6.3					0,000		
7	0,015	0,011	0,009	0,024	0,015	0,012	0,013
7.1		0,000					
7.3					0,000		
8	0,157	0,151	0,188	0,159	0,175	0,177	0,161
9	0,121	0,127	0,137	0,096	0,118	0,133	0,122
9.1					0,000		
9.3	0,000						
10	0,275	0,257	0,291	0,269	0,280	0,291	0,283
10.2	0,000						
11	0,236	0,242	0,214	0,245	0,226	0,228	0,232
12	0,161	0,169	0,137	0,173	0,154	0,129	0,150
12.3		0,000			0,000		
13	0,030	0,037	0,021	0,029	0,028	0,024	0,036
14	0,004	0,005		0,005	0,003	0,007	0,003
15	0,000				0,000		
17	0,000						

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP1: São Paulo (Fridman et al. 2008); RP: Ribeirão Preto (Presente estudo).

Tabela C8. Frequências alélicas no *locus* D8S1179.

Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
3	0,000						
4	0,000						
5	0,000						
6	0,000						
7	0,000						
8	0,008	0,006		0,005	0,005	0,008	0,006
9	0,009	0,008	0,005		0,007	0,011	0,016
10	0,068	0,078	0,038	0,098	0,061	0,073	0,082
11	0,077	0,065	0,052	0,083	0,076	0,082	0,084
12	0,130	0,143		0,123	0,124	0,126	0,140
13	0,279	0,280	0,267	0,250	0,276	0,289	0,247
14	0,251	0,238	0,295	0,270	0,258	0,244	0,249
15	0,138	0,145	0,176	0,127	0,148	0,134	0,134
16	0,034	0,034	0,057	0,034	0,038	0,027	0,032
17	0,004	0,004	0,010	0,010	0,005	0,006	0,009
18	0,001	0,000			0,000		0,001

							<i>Conclusão</i>
Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
19	0,000						
20	0,000						
21	0,000						
22	0,000						

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C9 Frequências alélicas no *locus* D10S1248.

Alelos	BR²	SP²	RP
7	0,001		
8	0,000		
9	0,002		0,001
10	0,001	0,002	0,001
11	0,019	0,010	0,009
12	0,053	0,059	0,061
13	0,271	0,281	0,269
14	0,308	0,309	0,328
15	0,206	0,189	0,182
16	0,109	0,114	0,125
17	0,028	0,033	0,019
18	0,002	0,002	0,003

BR2: Brasil (Moyses et al. 2017); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo)

Tabela C10. Frequências alélicas no *locus* D12S391.

Alelos	BR²	RP
15	0,040	0,055
15.1	0,000	
16	0,036	0,021
16.3	0,000	
17	0,104	0,122
17.1	0,002	
17.3	0,015	0,015
18	0,188	0,193
18.3	0,019	0,024
19	0,159	0,147
19.1	0,002	0,001
19.3	0,009	0,006
20	0,148	0,113
20.1		0,001
20.3	0,002	0,003
21	0,091	0,101
21.3	0,000	
22	0,089	0,095
23	0,059	0,055
24	0,019	0,036
25	0,014	0,009
26	0,003	0,001

Alelos	BR ²	RP
27	0,000	

BR2: Brasil (Moyses et al. 2017); RP: Ribeirão Preto (Presente estudo)

Tabela C11. Frequências alélicas no *locus* D13S317.

Alelos	BR ¹	RJ	MS	MG	SP ²	RP
4	0,000					
5	0,000					
6	0,000					
7	0,001			0,001	0,002	
8	0,100	0,073	0,077	0,095	0,088	0,125
9	0,082	0,056	0,091	0,068	0,077	0,088
10	0,053	0,030	0,024	0,045	0,051	0,058
11	0,295	0,316	0,322	0,305	0,279	0,305
12	0,293	0,329	0,308	0,309	0,325	0,262
13	0,126	0,137	0,130	0,130	0,130	0,119
14	0,048	0,060	0,048	0,048	0,048	0,042
15	0,001			0,001		0,001
16	0,000					

BR1: Brasil (Aguiar et al. 2012); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014) (Fridman et al. 2008); RP: Ribeirão Preto (Presente estudo)

Tabela C12. Frequências alélicas no *locus* D16S539.

Alelos	BR ¹	BR ²	RJ	MS	MG	SP	RP
2	0,000						
5	0,001	0,000			0,000	0,000	
6	0,000				0,000		
7	0,000						0,001
8	0,024	0,016	0,047	0,029	0,024	0,020	0,026
9	0,160	0,175	0,235	0,163	0,164	0,166	0,138
9.2	0,000						
10	0,088	0,095	0,094	0,053	0,084	0,098	0,096
10.2	0,000						
11	0,289	0,270	0,261	0,298	0,295	0,279	0,303
12	0,256	0,269	0,214	0,260	0,252	0,260	0,243
13	0,155	0,147	0,132	0,149	0,155	0,154	0,163
14	0,025	0,027	0,017	0,048	0,023	0,021	0,024
15	0,002				0,001	0,000	0,006
16	0,000						
17	0,000						

BR1: Brasil (Aguiar et al. 2012); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014) (Fridman et al. 2008); RP: Ribeirão Preto (Presente estudo)

Tabela C13. Frequências alélicas no *locus* D18S51.

Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
4	0,000						
5	0,000						
6	0,000						
7	0,000						
8	0,000						
9	0,001				0,001	0,000	0,001
10	0,010	0,012	0,005	0,010	0,012	0,007	0,010
10.2	0,001	0,001		0,015	0,001		
11	0,011	0,009	0,014	0,015	0,009	0,009	0,010
12	0,121	0,115	0,038	0,093	0,118	0,127	0,138
12.2	0,000				0,000		
12.3	0,000						
13	0,109	0,101	0,086	0,108	0,099	0,115	0,121
13.2	0,005	0,002	0,005		0,001	0,000	
13.3	0,000						
14	0,147	0,166	0,124	0,176	0,128	0,151	0,145
14.2	0,000	0,001	0,005		0,001		
14.3	0,000						
14.5	0,000						
15	0,152	0,168	0,133	0,147	0,147	0,152	0,124
15.2	0,000	0,000			0,000		
16	0,144	0,139	0,200	0,132	0,156	0,133	0,136
16.2	0,000					0,000	
16.5	0,000						
17	0,122	0,116	0,100	0,132	0,125	0,121	0,135
18	0,078	0,074	0,119	0,108	0,083	0,081	0,065
18.2	0,000		0,005				
19	0,049	0,049	0,086	0,049	0,060	0,049	0,064
19.2	0,000						
20	0,029	0,025	0,057	0,015	0,036	0,031	0,031
20.2	0,000						
21	0,013	0,015	0,014	0,015	0,014	0,012	0,010
21.2	0,000	0,000			0,000	0,001	
22	0,006	0,003	0,010		0,007	0,006	0,004
22.2					0,000		
23	0,001	0,002			0,001	0,001	
23.2	0,000						
24	0,001	0,001			0,001	0,001	0,003
25	0,000				0,000		0,001
26	0,000	0,000				0,000	
27	0,000	0,000			0,000	0,000	
28	0,000					0,001	

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo)

Tabela C14. Frequências alélicas no *locus* D19S433.

Alelos	BR2	MG	SP2	RP
5.2	0,001			
9		0,000	0,000	
10	0,002	0,004	0,001	0,004
10.2	0,000	0,000		
11	0,014	0,031	0,308	0,022
11.2	0,001	0,000	0,000	
12	0,100	0,103	0,102	0,108
12.1	0,000			
12.2	0,015	0,016	0,010	0,009
13	0,264	0,249	0,254	0,246
13.2	0,050	0,034	0,027	0,025
14	0,266	0,275	0,301	0,293
14.2	0,036	0,042	0,032	0,050
15	0,147	0,140	0,139	0,143
15.2	0,057	0,053	0,061	0,049
16	0,026	0,033	0,034	0,031
16.2	0,016	0,017	0,016	0,013
17	0,003	0,002	0,003	0,003
17.2	0,001	0,001	0,002	0,003
18.2		0,000	0,000	

BR2: Brasil (Moyses et al. 2017); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo)

Tabela C15. Frequências alélicas no *locus* D21S11.

<i>Continua</i>							
Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
12	0,000						
22	0,000						
23	0,000						
24	0,000						0,003
24.2	0,002	0,002			0,002	0,002	
24.3		0,003				0,001	
25	0,001	0,002			0,001		
25.2	0,001	0,002			0,001	0,001	
26	0,001	0,001			0,001		0,002
26.2	0,000						
27	0,030	0,029	0,057	0,025	0,039	0,026	0,035
27.2	0,000						
28	0,159	0,158	0,171	0,113	0,175	0,157	0,155
28.2	0,003				0,000		
28.3		0,001					0,005
29	0,208	0,191	0,205	0,211	0,215	0,219	0,223
29.1		0,000					
29.2	0,005	0,001			0,001		
29.3		0,000					0,005
30	0,232	0,247	0,243	0,240	0,220	0,233	0,224
30.2	0,028	0,030	0,024	0,039	0,030	0,027	0,036
31	0,077	0,069	0,052	0,103	0,067	0,069	0,081
31.2	0,090	0,105	0,086	0,098	0,097	0,103	0,092
32	0,028	0,009	0,005	0,015	0,013	0,011	0,003
32.2	0,075	0,098		0,078	0,085	0,091	0,080

							Conclusão
Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
33	0,012				0,005	0,005	0,002
33.1		0,000				0,001	
33.2	0,028	0,034	0,057	0,059	0,031	0,037	0,036
33.3	0,000						
34	0,005	0,001	0,005		0,004	0,003	0,006
34.2	0,004	0,006		0,005	0,002	0,002	0,003
35	0,006	0,007	0,024	0,005	0,009	0,006	0,011
35.1	0,000	0,000				0,001	
35.2	0,001			0,005	0,000	0,000	
36	0,002	0,001	0,005	0,005	0,003	0,003	
36.2	0,000					0,000	
37	0,001	0,001			0,000	0,001	
38	0,000				0,000		
39	0,000	0,001					

BR1: Brasil (Aguilar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C16. Frequências alélicas no *locus* D22S1045.

Alelos	BR2	SP2	RP
8		0,002	
9	0,000		0,001
10	0,009	0,009	0,018
11	0,103	0,121	0,106
12	0,018	0,017	0,015
13	0,004	0,005	0,006
14	0,036	0,042	0,051
15	0,356	0,359	0,358
16	0,350	0,330	0,310
17	0,109	0,109	0,129
18	0,012	0,007	0,006
19	0,003	0,000	

BR2: Brasil (Moyses et al. 2017); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C17. Frequências alélicas no *locus* F13A01.

					Continua
Alelos	DF	RS	RJ (EUR)	RP	
3.2	0,161	0,122	0,077	0,116	
4	0,060	0,063	0,114	0,055	
5	0,227	0,230	0,191	0,196	
6	0,214	0,246	0,283	0,252	
7	0,252	0,282	0,260	0,307	
8	0,023	0,019	0,029	0,019	
9	0,004		0,006		
10	0,004	0,001			
11	0,002	0,002	0,003	0,004	
12	0,006	0,002	0,003	0,003	
13	0,009	0,007	0,011	0,013	
14	0,011	0,005	0,009	0,006	

<i>Conclusão</i>				
Alelos	DF	RS	RJ (EUR)	RP
15	0,020	0,015	0,014	0,021
16	0,009	0,008	0,003	0,004
17	0,001			0,003

DF: Distrito Federal (Dalton et al. 2009); RS: Rio Grande do Sul (Gastaldo et al. 2015); RJ (EUR): Indivíduos de ancestralidade europeia do Rio de Janeiro (Silva e Moura-Neto 2004); RP: Ribeirão Preto (Presente estudo).

Tabela C18. Frequências alélicas no *locus* F13B.

Alelos	DF	RS	RJ (EUR)	RP
6	0,194	0,133	0,132	0,169
7	0,062	0,032	0,026	0,031
8	0,162	0,212	0,209	0,209
8.2		0,001		
9	0,260	0,249	0,255	0,218
10	0,320	0,370	0,364	0,370
11	0,002	0,003	0,007	0,003
12		0,001	0,013	

DF: Distrito Federal (Dalton et al. 2009); RS: Rio Grande do Sul (Gastaldo et al. 2015); RJ (EUR): Indivíduos de ancestralidade europeia do Rio de Janeiro (Silva e Moura-Neto 2004); RP: Ribeirão Preto (Presente estudo).

Tabela C19. Frequências alélicas no *locus* FES/FPS.

Alelos	DF	RS	RJ (EUR)	RP
7		0,001	0,006	0,032
8	0,041	0,022	0,018	0,003
9	0,012	0,006	0,023	
10	0,261	0,248	0,284	0,271
10.3		0,005		
11	0,415	0,429	0,365	0,416
11.3		0,001		
12	0,218	0,237	0,266	0,214
13	0,047	0,047	0,044	0,056
14	0,006	0,005	0,003	0,003

DF: Distrito Federal (Dalton et al. 2009); RS: Rio Grande do Sul (Gastaldo et al. 2015); RJ (EUR): Indivíduos de ancestralidade europeia do Rio de Janeiro (Silva e Moura-Neto 2004); RP: Ribeirão Preto (Presente estudo).

Tabela C20. Frequências alélicas no *locus* FGA

Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
7	0,000						
8	0,000						
9	0,000						
10	0,000						
11	0,000						
12	0,000		0,110				
13	0,000						
14	0,000						

Continua

Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
15	0,000						
16	0,001				0,000		
17	0,002	0,001	0,005		0,001	0,002	0,002
17.2	0,000				0,000		
18	0,009	0,010	0,010	0,015	0,008	0,009	0,014
18.2	0,002	0,003			0,005	0,003	0,009
18.3	0,000						
19	0,075	0,076	0,067	0,083	0,074	0,067	0,078
19.2	0,001	0,000			0,001	0,000	0,002
20	0,114	0,105	0,062	0,118	0,116	0,121	0,129
20.2	0,000	0,000	0,005		0,000	0,000	0,006
21	0,151	0,131	0,171	0,167	0,146	0,161	0,149
21.2	0,001	0,002			0,002	0,001	0,003
22	0,162	0,160	0,157	0,113	0,169	0,166	0,150
22.2	0,002	0,005			0,003	0,003	0,006
22.3		0,000				0,000	
23	0,149	0,140	0,162	0,181	0,148	0,151	0,135
23.1						0,000	
23.2	0,001	0,000		0,005	0,001	0,005	
24	0,151	0,161	0,143	0,123	0,154	0,150	0,138
24.2	0,000	0,000			0,000	0,000	
24.3		0,001					0,002
25	0,103	0,114	0,100	0,132	0,099	0,099	0,108
25.2	0,000	0,000			0,000		
26	0,047	0,054	0,048	0,049	0,045	0,000	0,041
26.2	0,000				0,000		
27	0,016	0,019	0,019	0,005	0,016	0,007	0,014
27.1							0,002
28	0,005	0,006	0,029		0,006	0,003	0,009
28.2	0,000						
29	0,002	0,004			0,002	0,002	0,002
29.2	0,000		0,010		0,000		
30	0,001	0,000	0,005	0,005	0,000		
30.2	0,000				0,000		
31	0,000		0,005				
31.2	0,001	0,001			0,001		0,003
32	0,000			0,005			
32.2	0,000	0,001			0,000		
33	0,000						
33.2	0,000						
39.2	0,000						
40	0,000						
41.2	0,000				0,000		
42	0,000						
42.2	0,000						
43	0,000						
43.2	0,001	0,000			0,000		
44.2	0,000	0,000				0,000	
45.2	0,000				0,000	0,001	
46.2	0,000	0,000			0,000		
47	0,000						
47.2	0,000	0,000					
48.2	0,000						

Conclusão

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C21. Frequências alélicas no *locus* LPL.

Alelos	DF	RS	RJ (EUR)	RP
7	0,007	0,004		0,002
8	0,003	0,001	0,017	
9	0,058	0,048	0,086	0,071
10	0,439	0,443	0,310	0,449
11	0,225	0,239	0,362	0,272
12	0,208	0,217	0,207	0,171
13	0,055	0,043	0,034	0,035
14	0,005	0,005		

DF: Distrito Federal (Dalton et al. 2009); RS: Rio Grande do Sul (Gastaldo et al. 2015); RJ (EUR): Indivíduos de ancestralidade europeia do Rio de Janeiro (Silva e Moura-Neto 2004); RP: Ribeirão Preto (Presente estudo).

Tabela C22. Frequências alélicas no *locus* PENTA C.

Alelos	RS	RP
5	0,015	0,015
7	0,005	0,007
8	0,031	0,028
9	0,182	0,156
10	0,053	0,055
11	0,334	0,364
12	0,249	0,218
13	0,115	0,135
14	0,016	0,019
15	0,001	0,003
16	0,001	

RS: Rio Grande do Sul (Gastaldo et al. 2015); RP: Ribeirão Preto (Presente estudo).

Tabela C23. Frequências alélicas no *locus* PENTA D.

Alelos	BR1	RS	RP
1	0,000		
2.2	0,033	0,017	0,043
2.3	0,000		
3	0,000		
3.2	0,003	0,001	
5	0,013	0,006	0,015
5.2	0,000		
6	0,001	0,002	0,002
6.4		0,001	
7	0,014	0,005	0,011
8	0,051	0,031	0,044
9	0,184	0,218	0,195
9.1		0,001	

Continua

Alelos	BR1	RS	RP
9.2	0,000	0,003	
9.3	0,000		
10	0,141	0,137	0,125
10.2	0,000		
11	0,157	0,149	0,155
11.2	0,000		
12	0,165	0,182	0,147
12.2	0,000		
13	0,154	0,163	0,170
14	0,060	0,066	0,071
14.3			0,002
15	0,019	0,016	0,020
16	0,003	0,002	0,002
17	0,001	0,001	
18	0,000		
19	0,000		
22	0,000		

BR1: Brasil (Aguiar et al. 2012); RS: Rio Grande do Sul (Gastaldo et al. 2015); RP: Ribeirão Preto (Presente estudo).

Tabela C24. Frequências alélicas no *locus* PENTA E.

Alelos	BR1	RS	RJ	MS	DF	RP
<i>Continua</i>						
2	0,000					
3	0,000					
4	0,000					
5	0,061	0,067	0,071	0,083	0,067	0,098
6	0,008		0,000			
7	0,109	0,115	0,157	0,093	0,106	0,171
8	0,058	0,038	0,081	0,025	0,060	0,082
9	0,023	0,016	0,024	0,015	0,019	0,024
9.3	0,000					
10	0,072	0,074	0,048	0,078	0,076	0,071
11	0,110	0,115	0,076	0,113	0,109	0,090
12	0,168	0,168	0,152	0,221	0,156	0,136
12.2	0,000					
13	0,113	0,101	0,095	0,103	0,119	0,103
14	0,069	0,065	0,100	0,074	0,079	0,068
15	0,062	0,069	0,048	0,029	0,055	0,054
15.1	0,000					
16	0,048	0,044	0,048	0,049	0,056	0,038
16.2	0,000					
17	0,040	0,046	0,052	0,039	0,035	0,030
18	0,022	0,020	0,019	0,034	0,025	0,008
19	0,014	0,022	0,014	0,015	0,008	0,016
20	0,010	0,019	0,010	0,015	0,015	0,005
21	0,006	0,009	0,005	0,010	0,010	0,005
22	0,004	0,008			0,006	
23	0,001	0,003		0,005	0,002	
24	0,000					
25	0,000					

						Conclusão
Alelos	BR1	RS	RJ	MS	DF	RP
26	0,000	0,001				
27	0,000					
28	0,000					

BR1: Brasil (Aguiar et al. 2012); RS: Rio Grande do Sul (Gastaldo et al. 2015); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); DF; Distrito Federal (Dalton et al. 2009); Ribeirão Preto (Presente estudo).

Tabela C25. Frequências alélicas no *locus* SE33.

				Continua
Alelos	BR2	SP2	RP	
5.2		0,000		
8	0,000			
9	0,000			
9.2	0,000			
10	0,000		0,007	
10.2	0,000	0,001		
11	0,002	0,001		
11.2	0,002	0,002	0,007	
12	0,001	0,003		
12.2	0,007	0,001	0,007	
13	0,008	0,007	0,027	
13.2	0,002	0,003	0,014	
14	0,027	0,028	0,051	
14.2	0,002	0,003	0,003	
15	0,055	0,043	0,047	
15.2	0,003	0,002	0,003	
16	0,070	0,077	0,088	
16.2	0,001	0,001	0,007	
16.3	0,001	0,001		
17	0,070	0,074	0,074	
17.3	0,001	0,001		
18	0,099	0,099	0,128	
18.1	0,000			
18.2	0,001	0,000		
18.3	0,000	0,000	0,003	
19	0,090	0,078	0,088	
19.2	0,001	0,002	0,003	
19.3	0,000			
20	0,054	0,057	0,068	
20.2	0,005	0,005	0,007	
21	0,034	0,036	0,027	
21.2	0,010	0,011	0,014	
22	0,012	0,015	0,020	
22.2	0,024	0,020	0,020	
23	0,003	0,002	0,003	
23.2	0,024	0,026	0,037	
24.2	0,028	0,028	0,017	
25	0,001	0,000		
25.2	0,033	0,032	0,041	

Alelos	BR2	SP2	RP
26	0,002		
26.2	0,055		0,034
27.2	0,059	0,064	0,034
27.3		0,000	
28	0,000		
28.2	0,066	0,071	0,051
29.2	0,048	0,056	0,024
29.3			
30		0,000	
30.2	0,036	0,042	0,027
31	0,001		
31.2	0,024	0,027	0,014
32	0,002	0,000	
32.2	0,015	0,013	0,007
33	0,005	0,000	
33.2	0,005	0,002	
33.3		0,000	
34	0,002	0,001	
34.2	0,001	0,001	
35	0,000	0,002	
36	0,000	0,000	
37	0,000	0,000	
38	0,000		

BR2: Brasil (Moyses et al. 2017); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C26. Frequências alélicas no *locus* TH01.

Continua

Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
4	0,000				0,000		
5	0,002	0,000		0,005	0,001	0,002	
6	0,211	0,220	0,133	0,245	0,191	0,223	0,212
6.3		0,000					
7	0,240	0,257	0,310	0,225	0,251	0,213	0,219
8	0,149	0,151	0,267	0,162	0,164	0,139	0,124
8.2	0,000				0,000		
8.3		0,000					
9	0,163	0,143	0,152	0,172	0,169	0,178	0,179
9.2	0,000						
9.3	0,226	0,218	0,129	0,186	0,214	0,229	0,253
10	0,008	0,009	0,010	0,005	0,009	0,014	0,013
10.3		0,000			0,001	0,000	
11	0,000				0,000	0,002	
12	0,000						
13	0,000						
13.3	0,000						
14	0,000						
15	0,000						
16	0,000						
18	0,000						
29.2			0,205				

							<i>Conclusão</i>
Alelos	BR1	BR2	RJ	MS	MG	SP2	RP
32.2			0,006				

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Tabela C27. Frequências alélicas no *locus* TPOx.

Alelos	BR1	BR2	RJ	MS	MG	SP1	RP
4	0,000						
5	0,001	0,001			0,000		0,001
6	0,018	0,014	0,052	0,010	0,022	0,027	0,021
7	0,008	0,006	0,014	0,000	0,007	0,011	0,006
8	0,456	0,466	0,400	0,436	0,449	0,461	0,449
9	0,124	0,106	0,162	0,093	0,138	0,135	0,140
10	0,067	0,064	0,062	0,083	0,068	0,063	0,063
11	0,276	0,285	0,257	0,358	0,278	0,263	0,272
12	0,048	0,056	0,052	0,015	0,035	0,038	0,047
13	0,002	0,002		0,005	0,001	0,002	0,001
14	0,000						
15	0,000						
16	0,000						
17	0,000						
18	0,000						

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP1: São Paulo (Fridman et al. 2008); RP: Ribeirão Preto (Presente estudo).

Tabela C28. Frequências alélicas no *locus* vWA.

							<i>Continua</i>
Alelos	BR2	BR2	RJ	MS	MG	SP2	RP
8	0,000						
10	0,000						
11	0,004	0,003	0,019		0,004	0,003	0,003
12	0,001	0,000			0,000		
13	0,006	0,003	0,014	0,010	0,007	0,003	0,006
14	0,084	0,076	0,095	0,088	0,083	0,088	0,086
14.1						0,000	
15	0,138	0,135	0,176	0,113	0,150	0,123	0,120
15.2	0,000						
16	0,260	0,276	0,290	0,275	0,253	0,259	0,235
16.2	0,000						
17	0,255	0,274	0,252	0,275	0,249	0,263	0,273
17.2	0,000						
17.3	0,000					0,000	
18	0,170	0,153	0,119	0,142	0,171	0,168	0,179
19	0,066	0,062	0,024	0,074	0,067	0,077	0,080
20	0,014	0,016	0,010	0,020	0,014	0,013	0,015

							Conclusão
Alelos	BR2	BR2	RJ	MS	MG	SP2	RP
21	0,002	0,002		0,005	0,003	0,003	0,003
22	0,000				0,000		
23	0,000						

BR1: Brasil (Aguiar et al. 2012); BR2: Brasil (Moyses et al. 2017); RJ: Rio de Janeiro (Silva et al. 2004); MS: Mato Grosso do Sul (Silva et al. 2004); MG: Minas Gerais (del Castillo et al. 2009); SP2: São Paulo (Almeida Prado Oliveira e Sousa et al. 2014); RP: Ribeirão Preto (Presente estudo).

Referências

- Aguiar VR, Wolfgramm Ede V, Malta FS, Bosque AG, Mafia Ade C, Almeida VC, Caxito Fde A, Pardini VC, Ferreira AC, e Louro ID. 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(4):504-509.
- Almeida Prado Oliveira e Sousa ML, de Oliveira MA, Auler-Bittencout EA, Soares-Vieira JA, Munoz DR, e Iwamura ES. 2014. Population data of 16 autosomal STR loci of the Powerplex ESX 17 System in a Brazilian Population from the State of Sao Paulo. *Forensic Sci Int Genet* 11:e15-17.
- Dalton GC, Maia FAS, Mendes CRBO, Pak HI, de Paula KAA, Trindade-Filho A, e Oliveira SF. 2009. Genetic profile of Federal District of Brazil based on 18 STR autosomal loci. *Forensic Science International: Genetics Supplement Series* 2(1):373-375.
- del Castillo DM, Perone C, de Queiroz AR, Mourao PH, de Souza Vasconcellos L, do Nascimento MA, and Januario JN. 2009. Populational genetic data for 15 STR markers in the Brazilian population of Minas Gerais. *Leg Med (Tokyo)* 11(1):45-47.
- Fridman C, dos Santos PC, Kohler P, Garcia CF, Lopez LF, Massad E, e Gattas GJ. 2008. Brazilian population profile of 15 STR markers. *Forensic Sci Int Genet* 2(2):e1-4.
- Gastaldo AZ, Rodenbusch R, Fre NND, Mardini AC, e Alho CS. 2015. Allele frequencies of 20 autosomal STR in the population from Rio Grande do Sul Southern Brazil. *Forensic Sci Int Genet* 19:15-17.
- Moyses CB, Tsutsumida WM, Raimann PE, da Motta CH, Nogueira TL, Dos Santos OC, de Figueiredo BB, Mishima TF, Candido IM, de Oliveira Godinho NM et al. . 2017.

Population data of the 21 autosomal STRs included in the GlobalFiler((R)) kits in population samples from five Brazilian regions. *Forensic Sci Int Genet* 26:e28-e30.

Silva DA, Crouse CA, Chakraborty R, Goes AC, e Carvalho EF. 2004. Statistical analyses of 14 short tandem repeat loci in Brazilian populations from Rio de Janeiro and Mato Grosso do Sul states for forensic and identity testing purposes. *Forensic Sci Int* 139(2-3):173-176.

Silva R, e Moura-Neto RS. 2004. Genetic diversity and admixture data on 11 STRs (F13B, TPOX, CSF1PO, F13A01, D7S820, LPL, TH01, vWA, D13S317, FESFPS, and D16S539) in a sample of Rio de Janeiro European-descendants population, Brazil. *Forensic Sci Int* 142(1):51-53.

Anexo A – Termo de Consentimento Livre e Esclarecido (TCLE)

Nome da pesquisa: Sequenciamento de nova geração de genes envolvidos na biossíntese de melanina em amostra da população brasileira

Pesquisador responsável: Prof. Dr. Celso Teixeira Mendes Junior

Endereço para contato: Av. Bandeirantes, 3900; CEP: 14040-901, Ribeirão Preto - SP
Departamento de Química, FFCLRP-USP Telefone: (16) 3602-0417

Prezado (a) doador (a),

A pigmentação humana é determinada pela presença de um pigmento denominado melanina. Proteínas produzidas por mais de 120 genes estão envolvidas na produção de melanina em células específicas e na distribuição desta melanina para outras células de nosso organismo. Variações normais nos genes envolvidos nestes processos levam aos diferentes tons de cor de cabelos, olhos e pele que vemos nas pessoas.

Nesse estudo, faremos a análise da sequência de algumas regiões do DNA que podem estar relacionadas à determinação das características físicas dos brasileiros, como a cor de olhos, pele e cabelos, tipo de cabelos e presença de sardas. Com isso buscamos identificar as variações (polimorfismos do DNA) e os mecanismos responsáveis pela determinação da cor das pessoas. Embora o conhecimento a ser produzido neste estudo não lhe trará benefícios diretos, os resultados desta pesquisa podem trazer grande contribuição para o desenvolvimento de terapias para tratar problemas de pele.

Para a realização desta pesquisa não haverá a necessidade de procedimentos que possam comprometer sua saúde. Caso concorde em participar da pesquisa, serão coletados cerca de 10 mL de seu sangue (aproximadamente uma colher de sopa). A coleta será feita por um profissional habilitado e altamente capacitado nessa atividade. Serão ainda obtidas imagens fotográficas de regiões pequenas e bem delimitadas de seu corpo (olho, raiz dos cabelos e regiões do braço não-expostas ao sol), as quais não permitirão sua identificação, mantendo, assim, o caráter confidencial da informação. Adicionalmente, um colorímetro portátil, equipamento frequentemente usado em pesquisas dermatológicas, será usado para definir a cor de sua pele e dos seus cabelos com maior precisão, em um procedimento rápido indolor e livre de qualquer desconforto. Em todo este procedimento não existem riscos previsíveis além de um pequeno desconforto durante a coleta de sangue.

Os seguintes aspectos lhe são esclarecidos para que você decida sobre sua participação voluntária:

- a) Esta pesquisa não lhe trará qualquer custo. Visando lhe proporcionar maior conforto e comodidade, a coleta será realizada uma única vez em data e local a combinar.
- b) Não haverá nenhuma forma de reembolso em dinheiro, uma vez que sua participação na pesquisa não proporcionará nenhum gasto.
- c) Você terá a garantia de receber a resposta a qualquer pergunta ou esclarecimento a qualquer dúvida acerca dos procedimentos, riscos, benefícios e outros relacionados com a pesquisa. Isso poderá ser feito a qualquer momento. Nós temos o compromisso de proporcionar informações atualizadas

durante o estudo, ainda que esta possa influenciar a sua vontade de continuar autorizando a sua participação.

- d) Você terá a liberdade de retirar o seu consentimento a qualquer momento sem que isso lhe traga prejuízo algum ou qualquer tipo de constrangimento.
- e) Você terá garantida a segurança de que não será identificado e que será mantida a sua privacidade e o caráter confidencial das informações concedidas.
- f) A amostra biológica será coletada com total responsabilidade e custo dos pesquisadores. A coleta de seu sangue será única e feita com material descartável, sendo necessário apenas um pequeno furo com agulha em veia do seu braço, o que pode, de fato, ser um procedimento desconfortável, mas apenas pela “picada” com a agulha.
- g) O sangue doado por você não será empregado na realização de pesquisas de outra natureza sem sua autorização.
- h) Será aplicado um questionário visando uma melhor compreensão de sua origem e características físicas. Toda informação obtida será mantida em sigilo.
- i) Os resultados da pesquisa serão apresentados em reuniões científicas (simpósios e congressos) e serão publicados em revistas científicas especializadas.

Assinatura do pesquisador responsável

Eu, _____, RG _____, residente _____, n° _____, cidade de _____, telefone _____ abaixo assinado, tendo recebido as informações acima e ciente dos meus direitos, concordo em participar da pesquisa por livre e espontânea vontade. Declaro ainda que recebi uma cópia deste termo de consentimento assinada pelo pesquisador responsável.

Tendo ciência do exposto acima, assino abaixo.

Ribeirão Preto, ____ de _____ de 20____.

Assinatura do doador

Anexo B - Termo de Consentimento Livre e Esclarecido para armazenamento de material biológico

Nome da pesquisa: Sequenciamento de nova geração de genes envolvidos na biossíntese de melanina em amostra da população brasileira

Pesquisador responsável: Prof. Dr. Celso Teixeira Mendes Junior

Endereço para contato: Av. Bandeirantes, 3900; CEP: 14040-901, Ribeirão Preto - SP
Departamento de Química, FFCLRP-USP Telefone: (16) 3602-0417

Prezado (a) doador (a),

Como você concordou em participar da pesquisa “Sequenciamento de nova geração das regiões regulatórias de dez genes envolvidos na biossíntese de melanina em amostra da população brasileira”, solicito agora uma nova autorização para que possamos armazenar parte do material biológico processado, para uso em pesquisas futuras desta mesma natureza, isto é, envolvendo a análise de outros genes e marcadores genéticos (polimorfismos do DNA).

Caso concorde, seu material biológico será armazenado sob minha responsabilidade em um freezer localizado em meu laboratório, compondo Biorrepositório devidamente registrado junto ao Comitê de Ética em Pesquisa desta Instituição. Este Biorrepositório será futuramente transformado em Biobanco, conforme determinação do Conselho Nacional em Saúde. Os seguintes aspectos lhe serão esclarecidos para que você decida sobre esta autorização voluntária:

- a) Você tem a liberdade de decidir se seu material será armazenado.
- b) Você terá a liberdade de retirar o seu consentimento a qualquer momento sem que isso lhe traga prejuízo algum ou qualquer tipo de constrangimento.
- c) O armazenamento não lhe trará qualquer custo. A amostra biológica será armazenada com total responsabilidade e custo dos pesquisadores.
- d) Toda nova pesquisa a ser realizada com o material armazenado será submetida para aprovação do Comitê de Ética em Pesquisa (CEP) institucional e, quando for o caso, da Comissão Nacional de Ética em Pesquisa (CONEP).
- e) Caso deseje, você tem o direito de ser consultado sempre que houver intenção de uso de seu material biológico em pesquisas futuras, para que eu autorize ou não autorize sua nova utilização.
- f) Você terá a garantia de receber a resposta a qualquer pergunta ou esclarecimento a qualquer dúvida acerca do seu material biológico, condições de armazenamento e sua utilização em outros projetos de pesquisa. Isso poderá ser feito a qualquer momento. Nós temos o compromisso de proporcionar informações atualizadas durante o estudo, ainda que esta possa influenciar a sua vontade de continuar autorizando a sua participação.

- g) Você terá garantida a segurança de que não será identificado e que será mantida a sua privacidade e o caráter confidencial das informações concedidas. Toda informação obtida será mantida em sigilo.
- h) Você tem o direito de ser informado sobre eventual perda ou destruição de suas amostras biológicas, bem como sobre eventual encerramento do Biorrepositório.
- i) Os resultados de novas pesquisas serão apresentados em reuniões científicas (simpósios e congressos) e serão publicados em revistas científicas especializadas.

Assinatura do pesquisador responsável

Eu, _____, RG no _____, residente na _____ n° _____, cidade de _____, telefone _____ abaixo assinado, tendo recebido as informações acima e ciente dos meus direitos, concordo com o armazenamento de meu material biológico por livre e espontânea vontade.

Declaro ainda que:

- () desejo exercer o direito de ser consultado sempre que houver intenção de uso de meu material biológico em pesquisas futuras.
- () dispenso o direito de ser consultado quando houver intenção de uso de meu material biológico em pesquisas futuras, autorizando automaticamente seu uso em estudos desta mesma natureza, isto é, envolvendo a análise de outros genes e marcadores genéticos (polimorfismos do DNA).

Atesto que recebi uma cópia deste termo de consentimento assinada pelo pesquisador responsável. Tendo ciência do exposto acima, assino abaixo.

Ribeirão Preto, ____ de _____ de 20__.

Assinatura do doador

Anexo C – Protocolo de análise laboratorial (*HaloPlex*) para sequenciamento de nova geração

Laboratório de Pesquisas Forenses e Genômicas
Departamento de Química – FFCLRP/USP

Protocolo de análise laboratorial (*HaloPlex*) para sequenciamento de nova geração

Oliveira MLG, Marcorin L, Pereira ALE, Debortoli G,

Fracasso NCA, Silva GV, Mendes-Junior CT

Versão 1 – Outubro de 2016

Adaptado de: Oliveira MLG. Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira. Dissertação de Mestrado apresentada à Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. 2016. 151p.

<u>1. EXTRAÇÃO DE DNA GENÔMICO</u>	130
<u>2. QUANTIFICAÇÃO E VERIFICAÇÃO DA QUALIDADE DO DNA</u>	130
<u>3. PREPARO DOS REAGENTES PARA ENRIQUECIMENTO DE ALVO</u>	131
<u>4. ENRIQUECIMENTO DO ALVO E PREPARO DA BIBLIOTECA</u>	131
<u>4.1. Digestão e hibridação do DNA</u>	132
<u>4.2. Captura, ligação, eluição e PCR</u>	133
<u>4.3. Purificação da biblioteca</u>	134
<u>5. CONTROLE DE QUALIDADE E QUANTIFICAÇÃO DA BIBLIOTECA</u>	134
<u>6. SEQUENCIAMENTO</u>	135
<u>7. REFERÊNCIAS</u>	136

1. EXTRAÇÃO DE DNA GENÔMICO

Após a coleta, a extração do DNA deve ser realizada a partir das amostras de sangue total utilizando o protocolo de *salting-out* (Miller, Dykes, & Polesky, 1988) com modificações.

O sangue coletado e mantido em tubo *Vacutainer*® com EDTA é transferido para um tubo *falcon* graduado de 50mL, ao qual é adicionado tampão de Lise I gelado (Tris-HCl 0,01M pH 7,5; Sacarose 0,3M; MgCl₂ 0,005M; Triton 1%; H₂O qsp) até completar o volume de 45mL. De forma a lisar as hemácias, o conteúdo é delicadamente misturado por inversão e centrifugado durante 15 minutos a 5000rpm em centrífuga 5430 R (Eppendorf AG). A extração prossegue com o botão de células obtido, o qual deve ser ressuspensionado em 4,5mL de tampão de Lise II (NaCl 0,075M, EDTA 0,024M, H₂O qsp, pH 8), 125μL de SDS 10% e 1,1mL de perclorato de sódio (5M) e vortexado a temperatura ambiente por 10 segundos para rompimento dos glóbulos brancos. Para lisar as proteínas, são adicionados 2mL de NaCl 6M (saturado) ao tubo *falcon*, que é vortexado a temperatura ambiente por 15 segundos. A seguir, o conteúdo é centrifugado por 15 minutos a 5000rpm em centrífuga 5430 R (Eppendorf AG). O sobrenadante é transferido para um novo tubo *falcon* de 50mL, ao qual é adicionado 5mL de isopropanol absoluto a temperatura ambiente (Merck Millipore, Darmstadt, Alemanha) e homogeneizado por inversão até que o DNA se separe e fique suspenso na solução.

Finalmente, o DNA é retirado com o auxílio de uma pipeta estéril e transferido para um microtubo de 1,5mL contendo 1mL de etanol (Merck Millipore, Darmstadt, Alemanha) 70% gelado. O microtubo é então centrifugado por 5 minutos a 5000rpm em centrífuga MiniSpin Plus (Eppendorf AG), e o sobrenadante descartado a seguir. O microtubo é deixado em centrífuga a vácuo (Concentrator plus, Eppendorf AG) durante 5min para remoção completa do etanol. Por fim, o DNA aderido ao microtubo é dissolvido em 200μL de água MilliQ, mantido na geladeira por 1 semana e posteriormente armazenado a -20°C no banco de amostras do laboratório.

2. QUANTIFICAÇÃO E VERIFICAÇÃO DA QUALIDADE DO DNA

O *NanoDrop*® ND-1000 (*Thermo Fisher Scientific Inc.*) é utilizado para avaliar o grau de contaminação do material genômico por proteínas e para estimativa inicial da concentração de DNA. É importante que as amostras possuam uma razão OD 260/280 variando de 1,8 a 2,0. Além disso, é realizada a eletroforese em gel de agarose, a fim de verificar o grau de

degradação do material, considerando como indicativo de degradação qualquer banda abaixo de 2,5kb (*Agilent Technologies*).

Na sequência, as amostras são quantificadas através de fluorescência por meio do *Qubit™ dsDNA BR Assay* (*Life Technologies*). Para isso, o DNA genômico é diluído em água MilliQ em uma proporção de 1:10. Inicialmente, é preciso fornecer ao instrumento uma curva padrão. Para tanto, o corante (fluoróforo) para dsDNA é diluído no *buffer* do Qubit (1:200) e 10µL dos padrões (1 e 2) são diluídos em 200µL da solução anterior. Realizada a leitura dos padrões, uma curva padrão é estabelecida, a partir da qual é calculada a concentração da amostra original, com base no volume adicionado para leitura.

Conhecendo agora a concentração das amostras, essas são diluídas à concentração de 5ng/µL, no intuito de normalizar as amostras a uma concentração compatível para a montagem das bibliotecas de fragmentos.

3. PREPARO DOS REAGENTES PARA ENRIQUECIMENTO DE ALVO

Regiões alvo, incluindo a sequência promotora, CDS (*Coding DNA Sequence*), 5'UTR e 3'UTR de genes humanos, além de outras regiões de interesse, foram submetidas ao ensaio *in silico* de desenho de sondas para captura de DNA alvo utilizando o aplicativo *SureDesign* da *Agilent Technologies* (<https://earray.chem.agilent.com/suredesign/home.htm>). Para desenho das sondas, o *SureDesign* tomou como referência as coordenadas dos genes em questão junto ao *release* hg19/GRCh37 do genoma humano. Visando a cobertura de 284.367pb, um total de 12.131 sondas foi desenhado, abrangendo 99,63% de tal extensão. Devido à distribuição dos sítios de restrição utilizados posteriormente no preparo das bibliotecas, as sondas desenhadas capturam uma região de 488.658pb, que garante o sequenciamento parcial dos introns flanqueadores dos exons alvo.

4. ENRIQUECIMENTO DO ALVO E PREPARO DA BIBLIOTECA

Após a quantificação, cada amostra de DNA é processada individualmente, gerando uma biblioteca contendo todos os amplicons. A captura das regiões-alvo utiliza reagentes do kit de enriquecimento de alvo customizado *HaloPlex* (*Agilent Technologies*) e segue o protocolo sugerido pelo fabricante. Resumidamente, o protocolo consiste dos seguintes passos: 1) digestão do DNA genômico com enzimas de restrição; 2) hibridação dos fragmentos de DNA a sondas biotiniladas cujos finais são complementares a esses

fragmentos-alvo (durante esse passo os fragmentos são circularizados e índices, que irão identificar cada indivíduo durante o sequenciamento, são adicionados); 3) captura dos fragmentos-alvo utilizando *beads* magnéticas recobertas por estreptavidina; 4) ligação dos fragmentos circularizados; 5) amplificação por PCR dos fragmentos-alvo capturados para montagem da biblioteca.

4.1. Digestão e hibridação do DNA

A digestão das amostras é feita conforme indicações do fabricante, com algumas modificações (*HaloPlex Target Enrichment System For Illumina Sequencing, Agilent Technologies*, 2013). Um total de 225ng (45µL de uma solução 5ng/µL) de DNA genômico de cada amostra e do *Enrichment Control DNA* (ECD), no volume total de 45µL, é separado e reservado em gelo.

Digestão: Para o mix de restrição, preparado para lotes de 12 amostras (11 amostras de DNA genômico e outra de ECD), são misturados 476µL de *RE buffer* e 11,9µL de BSA e esse conteúdo é distribuído em 8 tubos. Em cada um dos 8 tubos é adicionado um par de enzimas de restrição (7µL de cada). Cada amostra é digerida pelos 8 pares de enzima separadamente. Para isso, cada amostra é distribuída em 8 tubos (5µL por tubo) onde são adicionados 5µL dos mixes contendo as enzimas de restrição resultando, para um total de 12 amostras, em 96 tubos de digestão que são incubados no termociclador (*SureCycler 8800 Thermal Cycler, Agilent Technologies*) a 37°C por 30 minutos.

Hibridação: Para o mix de hibridação são misturados 650µL da *Hybridization Solution* e 260µL das sondas *Haloplex* em um tubo e 70µL dessa solução é distribuída por amostra. Adicionalmente, são diluídos nessa solução 10µL de um dos índices disponíveis em cada um dos doze tubos (índices diferentes são utilizados para cada amostra, permitindo sua identificação após a mistura das amostras para o sequenciamento).

O DNA digerido de cada amostra (80µL somando todos os 8 tubos de digestão) é então transferido para os microtubos contendo o mix de hibridação (70µL) e seus índices correspondentes (10µL). Essas soluções são incubadas a 54°C durante 16 horas. O tempo de duração da hibridação é determinado de acordo com a densidade de sondas do ensaio. Segundo instruções do fabricante (*Haloplex, Agilent Technologies*), ensaios que atingem cobertura de entre 500kb - 5Mb exigem 16 horas de hibridação, enquanto que ensaios com cobertura de entre 1-500kb exigem apenas 3 horas de hibridação. Apesar de o presente ensaio

estar voltado para a cobertura de apenas 488,7kb, os ensaios iniciais demonstraram a necessidade de um tempo de hibridação de 16 horas.

4.2. Captura, ligação, eluição e PCR

Captura: A captura dos fragmentos hibridados é feita com o uso de 520µL de *beads* magnéticas recobertas por estreptavidina. Com o auxílio de uma placa magnética (*DynaMag™-2 Magnet, Life Technologies*), o sobrenadante da solução de *beads* é removido e substituído por 520µL de *Capture Solution*. Deste conteúdo, 40µL é adicionado a cada solução de DNA hibridizado (160µL) e após 15min de incubação em temperatura ambiente, os tubos contendo as soluções são colocados junto a uma placa magnética de 96 poços (*Agencourt SPRIPlate 96R, Beckman Coulter*) para a remoção do sobrenadante. A seguir, adiciona-se 100µL de *Wash Solution* por amostra, e essa solução é incubada no termociclador a 46°C por 10min. Após esse período, o sobrenadante é novamente removido.

Ligação: Procedendo com a ligação e circularização dos fragmentos alvo, adiciona-se 617,5µL de *Ligation Solution* e 32,5µL de DNA ligase e distribuí-se 50µL dessa solução a cada amostra hibridada e capturada. A solução é então incubada a 55°C por 10min.

Eluição: A seguir, lavamos o conteúdo de DNA ligado com 100µL de *SSC buffer*, retirando o sobrenadante ao final da etapa com o auxílio da placa magnética. Para eluir o DNA das *beads* adicionamos 25µL de NaOH (50mM) e encubamos em temperatura ambiente por 1min..

PCR: Com o auxílio da placa magnética, coletamos o DNA eluído num total de 20µL por amostra e o transferimos para tubos contendo mix de PCR. Tal mix para a reação de PCR é preparado com 209,3µL de água MilliQ, 130µL de 5X *Herculase II Reaction Buffer*, 5,2µL de dNTPs (100mM, 25mM para cada dNTP), 13µL de cada um dos *primers forward* e *reverse* (25µM), 6,5µL de ácido acético (2M) e 13µL da polimerase *Herculase II Fusion*, somando um total de 30µL por amostra. O volume final da reação é de 50µL por amostra. As condições de ciclagem da PCR seguem uma etapa inicial de desnaturação por 10 minutos a 98°C, seguida por 20 ciclos de 98°C por 30 segundos, 60°C por 30 segundos e 72°C por 1 minuto, e um passo final a 72°C por 10 minutos. Após o término da reação de PCR o produto é mantido a 8°C.

4.3. Purificação da biblioteca

Os produtos de PCR são então purificados utilizando *AMPure XP beads* (Beckman Coulter) de acordo com as instruções da *Agilent Technologies*. Um total de 40µL do produto de PCR é misturado a 100µL de *AMPure XP beads* mais 40µL de água MilliQ por amostra. A seguir, essa solução é colocada junto à placa magnética para remoção do sobrenadante e é lavada duas vezes com álcool 70%, que é completamente removido ao final. Então, adicionamos 40µL de Tris-HCl (pH = 8,0) para eluir o DNA capturado pelas *AMPure XP beads*. O produto final de 40µL é armazenado a -20 °C.

5. CONTROLE DE QUALIDADE E QUANTIFICAÇÃO DA BIBLIOTECA

A qualidade das bibliotecas é um dos principais determinantes para o sucesso da corrida de sequenciamento. Bibliotecas de fragmentos que não correspondem à distribuição de tamanho recomendada podem levar a uma baixa cobertura ou à falha de cobertura em determinadas porções das sequências alvo (Cher, 2011).

O controle de qualidade das bibliotecas é feito através do sistema de eletroforese capilar *2100 Bioanalyzer* (*Agilent Technologies*) utilizando um chip de alta-sensibilidade. A eletroforese é capaz de separar os fragmentos com base no tamanho, permitindo dessa forma avaliar a qualidade das bibliotecas geradas. É esperado para este ensaio que as bibliotecas apresentem um perfil de picos variando de 175 a 625pb, conforme previsto por um eletroferograma padrão (Figura 1). A partir da seleção manual da faixa de fragmentos gerados (175 a 625pb) é possível obter a concentração (em pg/µL e pmol/L) referente a este intervalo, sendo a molaridade utilizada para a quantificação de cada uma das bibliotecas para composição adequada do *pool* de bibliotecas para o sequenciamento. Adicionalmente, as bibliotecas individuais podem ser quantificadas por fluorescência (*Qubit™ dsDNA BR Assay*, *Life Technologies*) para composição adequada do *pool* de bibliotecas para o sequenciamento.

Figura 1: Perfil de picos esperado no eletroferograma para as bibliotecas de fragmento deste ensaio. Tal eletroferograma é produzido pela *Agilent Technologies* durante a avaliação do kit *Haloplex* customizado por meio de ensaio *in silico* de desenho de sondas para captura de DNA alvo utilizando o aplicativo *SureDesign*. Para leitura correta no *2100 Bioanalyzer* é crítico que os 13 picos previstos no *ladder* sejam identificados, que os picos do menor (35pb) e do maior (10380pb) marcador apresentem um tempo de corrida correspondente entre as amostras analisadas e o *ladder* e que as linhas de base sejam planas (Cher, 2011).

6. SEQUENCIAMENTO

Como descrito no item anterior, a partir das concentrações obtidas no *2100 Bioanalyzer* e *Qubit*, prepara-se soluções a 4nM de biblioteca, de maneira que cada biblioteca tenha a mesma representatividade. O *2100 Bioanalyzer* é utilizado também para quantificação deste *pool* final de bibliotecas. O *pool* final de bibliotecas pode ser adicionalmente quantificado por fluorescência (*Qubit™ dsDNA BR Assay*, *Life Technologies*) ou PCR em tempo real (*ViiA™ 7 Real-Time PCR System*, *Applied Biosystems*). Na PCR quantitativa em tempo real (qPCR), através da construção de uma curva padrão, é possível determinar a quantidade absoluta de DNA nas amostras a serem sequenciadas (*Applied Biosystems ViiA™ 7 Real-Time PCR System. User Guide. Life Technologies*, 2010, 262p).

A seguir, 5µL do *pool* de bibliotecas a 4nM são desnaturados pela adição de 5µL de NaOH (0,2M) e 990µL de *Hybridization Buffer* (HT1) para obtenção de uma biblioteca a 20pM, que é diluída com HT1 para que se chegasse à 600µL de uma solução de concentração 16pM. Essa solução é inserida no cartucho de sequenciamento do *MiSeq Personal Sequencer*

(*Preparing Libraries for Sequencing on the MiSeq*, California, USA, *Illumina*, Inc, 2013, 14p).

Antes de iniciar o sequenciamento é preciso carregar uma *SampleSheet* ou planilha de amostras (*.csv) contendo as informações necessárias para configurar a corrida, as quais correspondem à lista de amostras a serem sequenciadas e às sequências dos índices utilizados em cada amostra.

Durante o sequenciamento, as moléculas de DNA fita simples se ligam à superfície da *flow cell* por meio dos adaptadores presentes em suas extremidades. O sequenciamento ocorre a partir da síntese da fita complementar ao fragmento ligado na *flow cell* utilizando didesoxinucleotídeos marcados com fluoróforos específicos para cada base (A, T, C e G). O *software* Bustard relaciona então a fluorescência emitida após cada incorporação ao didesoxinucleotídeo correspondente com o *cluster* de fragmentos fixados à região que emitiu o sinal. No caso do sequenciamento usando reagentes *MiSeq Reagent Kit V3 (600 cycle)*, são realizados 301 ciclos em leituras do tipo *paired-end* (cada fragmento é sequenciado tanto na fita *forward* quanto na *reverse*) perfazendo um total de 2x301 ciclos de leitura, além de 8 ciclos referentes à leitura do índice incorporado em cada fragmento capturado, o que relaciona de maneira unívoca o fragmento a uma determinada amostra.

7. REFERÊNCIAS

- Cher, C. (Field A. S. (2011). *Bioanalyzer Applications for Next Gen Sequencing: Updates and Tips*, 1–49.
- Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, 16(3), 1215. <https://doi.org/10.1093/nar/16.3.1215>

Anexo D – Protocolo de processamento computacional de dados de sequenciamento de nova geração

Laboratório de Pesquisas Forenses e Genômicas
Departamento de Química – FFCLRP/USP

Protocolo de processamento computacional de dados de sequenciamento de nova geração

Oliveira MLG, Marcorin L, Pereira ALE, Debortoli G,

Fracasso NCA, Silva GV, Castelli EC, Mendes-Junior CT

Versão 2 – abril de 2018

Adaptado de: Oliveira MLG. Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira. Dissertação de Mestrado apresentada à Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo. 2016. 151p.

<u>1. ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO</u>	138
<u>1.1. Análise primária</u>	138
<u>1.2. Análise secundária</u>	139
<u>1.2.1. Avaliação da qualidade</u>	139
<u>1.2.2. Trimagem</u>	139
<u>1.2.3. Alinhamento ao genoma de referência</u>	141
<u>1.2.3.1. SAM</u>	143
<u>1.2.3.2. BAM</u>	144
<u>1.3. Análise terciária</u>	145
<u>1.3.1. Determinação de variantes</u>	145
<u>1.3.1.1. VCF</u>	149
<u>1.3.2. Remoção de genótipos de baixa qualidade</u>	151
<u>2. REFERÊNCIAS</u>	153

1. ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

O fluxograma de análises no NGS pode ser sumarizado em três fases principais: a fase primária, que consiste na geração dos dados no formato FastQ, a fase secundária, que compreende a aplicação de filtros de qualidade e o alinhamento a um genoma de referência e a fase terciária, que compreende o processo de determinação das variantes através de comparação das sequências alvo a um genoma de referência.

1.1. Análise primária

O processamento dos dados segue passos gerais pré-estabelecidos, de acordo com o objetivo pretendido, e tem início no *base calling* (ou a interpretação da imagem obtida a partir do sequenciador), quando são gerados os dados brutos do sequenciamento no formato *base call* (*.bcl). Estes dados são então disponibilizados através do ambiente de computação em nuvem da *Illumina*, *BaseSpace* (<https://basespace.illumina.com>), no formato FastQ (*.fastq.gz), o qual possui uma pontuação de qualidade associada (*Q-score*) a cada base sequenciada.

Essa pontuação de qualidade, também denominada *Phred-score* por ter sido introduzida como convenção pelo programa Phred (Ewing, Hillier, Wendl, & Green, 1998), se relaciona à probabilidade de erro na determinação de uma base. A medida se baseia na fórmula a seguir, onde dado a determinação de uma base X, a probabilidade de que X não seja verdadeira ($P(X)$) é expressa pela relação $Q(X) = -10 \log_{10} (P(X))$ ou $P(X) = 10^{-Q(X)/10}$. Assim, um Q30 indica probabilidade de ocorrência de um erro a cada 1.000 bases determinadas.

Cada arquivo FastQ é nomeado como no exemplo a seguir, onde PIG0001 representa o nome da amostra, S1 representa o número da amostra e indica neste exemplo que esta é a primeira amostra listada na *SampleSheet*, L001 indica o número da linha que a amostra ocupa na *flow cell*, sendo sempre o mesmo, já que a *flow cell* utilizada possui apenas uma linha, R1 e R2 indicam que o arquivo é composto por *reads forward* e *reverse*, respectivamente, enquanto o último segmento 001 sempre recebe essa denominação:

```
PIG0001_S1_L001_R1_001.fastq.gz  
PIG0001_S1_L001_R2_001.fastq.gz
```

1.2. Análise secundária

Essa segunda etapa consiste na aplicação de filtros de qualidade e posterior alinhamento dos fragmentos sequenciados a um genoma de referência.

1.2.1. Avaliação da qualidade

Em um primeiro momento, é importante verificar a qualidade das *reads* sequenciadas, a fim de garantir que os dados não apresentem problemas que poderiam afetar o passo seguinte de mapeamento. Para tanto, pode ser empregado o *software* FastQC (Andrews S., 2010), o qual utiliza um arquivo FastQ para gerar um relatório completo da qualidade das sequências (Figura 1a). A partir dos resultados apresentados é possível decidir por estratégias que melhorem a qualidade dos dados gerados.

Figura 1: Gráfico de distribuição da qualidade por base no programa FastQC. Cada coluna apresenta um *boxplot* com os valores de qualidade média das bases incorporadas em todas as *reads* em um determinado ciclo. Como pode ser observado em (a) ocorre uma diminuição da qualidade ($Q < 30$) das *reads* não-processadas a partir dos ciclos 150-154. O eixo Y apresenta os *scores* de qualidade, sendo dividido conforme cores que indicam: verde (boa qualidade), laranja (qualidade razoável) e vermelho (baixa qualidade). (b) Após processadas pelo *cutadapt*, foi obtida uma maior proporção das *reads* com qualidade desejável, ainda que a partir dos ciclos 225-229 tenha ocorrido uma diminuição da qualidade, o que, no entanto, é comum para ensaios da *Illumina*.

1.2.2. Trimagem

Sequências de adaptadores, bem como bases diagnosticadas com baixo *Q-score* devem ser retiradas das extremidades das *reads*, em um processo referido como trimagem. As sequências obtidas podem ser trimadas de diversas maneiras: por tamanho, qualidade ou a partir de sequências conhecidas.

A trimagem por qualidade parte do princípio de que é preferível trabalhar com sequências menores de maior qualidade, removendo sequências de bases na extremidade da *read* que tenham uma qualidade abaixo do limite estabelecido como adequado. A partir do que pôde ser observado no relatório gerado pelo FastQC optou-se por realizar a trimagem apenas pelos adaptadores *paired-end TruSeq da Illumina* utilizados no sequenciamento.

No processo de sequenciamento *Illumina*, todos os *clusters* são sequenciados seguindo o número de ciclos previstos pelo kit de reagentes utilizado (no caso, 2x301 ciclos). Mesmo que um *cluster* inclua um pequeno fragmento a ser sequenciado, os ciclos de sequenciamento não são interrompidos, continuando até alcançar o número total de ciclos. Como consequência, as *reads* geradas como *output* podem conter, além das sequências alvo (caso essas sejam menores que o número de ciclos), as sequências dos adaptadores do outro extremo do fragmento. Caso ainda restem ciclos após o sequenciamento dos adaptadores, leituras adicionais continuam sendo feitas, embora com qualidade extremamente baixa, uma vez que nucleotídeos inexistentes são aleatoriamente identificados (possivelmente devido à fluorescência de diferentes nucleotídeos não incorporados e não retirados durante a lavagem). Logo, é preciso encontrar e remover esses adaptadores, o que resultará, consequentemente, na remoção das bases incorporadas aleatoriamente.

Em função da construção realizada pelo *Haloplex (PCR primer --- Illumina adaptor -- - TARGET --- Illumina adaptor --- Barcode --- PCR primer)*, ao remover as sequências dos adaptadores de uma *read* podemos remover tudo o que existe depois dessas sequências, neste caso os *primers* da PCR e o *barcode*, quando presente.

A trimagem pelos adaptadores pode ser realizada com uso do *cutadapt* (Martin, 2011) utilizando a seguinte linha de comando (Tabela 1):

```
cutadapt -e 0.2 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -A  
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -o  
sample.trimmed.r1.fastq -p sample.trimmed.r2.fastq  
input.raw.r1.fastq input.raw.r2.fastq
```

Tabela 1: Descrição dos componentes da linha de comando executada no *cutadapt*

Comandos utilizados no <i>cutadapt</i>	Função do comando executado
-e 0.2	indica o nível máximo de tolerância de erros (<i>mismatches</i> , inserções e deleções) entre as sequências dos adaptadores e sequências presentes nas <i>reads</i> , para que uma sequência adaptadora seja efetivamente trimada.
-a	indica que o parâmetro a seguir se refere à sequência do adaptador <i>forward</i> .
-A	indica que o parâmetro a seguir se refere à sequência do adaptador <i>reverse</i> .
-o	indica o arquivo <i>output</i> da primeira <i>read</i> do par.
-p	indica o arquivo <i>output</i> da segunda <i>read</i> do par.

Tal procedimento resultou em melhorias significativas na qualidade média das *reads* processadas (Figura 1b).

1.2.3. Alinhamento ao genoma de referência

Estando as *reads* processadas, o próximo passo é mapeá-las contra um genoma de referência. A maior parte dos programas de alinhamento disponíveis usam um dentre os seguintes algoritmos de mapeamento: baseado em *hash* ou na transformação de Burrows-Wheeler (BWT). Estes algoritmos irão construir um arquivo de alinhamento com base nos dados de *single-end* e em uma análise seguinte utilizarão os dados de *paired-end* para refinar os alinhamentos gerados.

Para fazer o alinhamento pode ser utilizado o BWA (Li & Durbin, 2009), o qual utiliza um algoritmo do tipo BWT, que tem como principal vantagem o menor consumo de memória. O BWA é um pacote de softwares, que consiste em três algoritmos, que são chamados por diferentes comandos: *aln/samse/sampe* para o BWA-backtrack, *bwasw* para o BWA-SW e *men* para o BWA-MEM.

Antes de realizar o alinhamento, porém, é preciso indexar o genoma de referência por meio do índice FM (Ferragina, Manzini, Veli, & Navarro, 2004). Neste caso, utilizamos um genoma já indexado (*Homo sapiens*, UCSC, hg19), disponibilizado pela *Illumina* através do *iGenomes* em: https://support.illumina.com/sequencing/sequencing_software/igenome.html. Pode-se também indexar localmente o genoma de interesse por meio de ferramentas específicas. Para tanto, é necessário gerar três diferentes arquivos, arquivo índice, arquivo

fasta índice e um dicionário de sequência, os quais podem ser obtidos por meio dos programas BWA, Samtools e Picard, respectivamente.

O algoritmo utilizado para realizar o alinhamento é o BWA-MEM em função da sua especificidade para sequências com tamanho entre 70 e 1Mb e por ser mais rápido e acurado que o BWA-SW. A seguinte linha de comando deve ser utilizada (Tabela 2):

```
bwa mem -t 6 -R
' \@R\G\\tID:sampleID\tLB:sampleLB\tSM:tag\tPL:illumina\tPU:sampleP '
/path/to/reference/hg19.fa
sample.trimmed.r1.fastq sample.trimmed.r2.fastq
> mapped.sample.sam
```

Tabela 2: Descrição dos componentes da linha de comando executada no BWA.

Comandos utilizados no BWA	Função do comando executado
-t	<i>number of threads</i> (número de processos com um endereço/espaco compartilhado no <i>Linux</i>).
-R	indica que o parâmetro a seguir se refere à linha do cabeçalho de determinado grupo de <i>reads</i> .
\t	comando convertido em TAB no output SAM, a fim de separar adequadamente os campos.
\@R\G\\	identifica o cabeçalho de cada grupo de <i>reads</i> . Permite ainda que o arquivo BAM possa ser lido pelo GATK.
ID	identifica a qual grupo de <i>reads</i> cada <i>read</i> pertence. O ID permite que ao invés de ter que lidar com múltiplos grupos de dados, se passe a lidar com apenas um. Assim, a informação sobre o grupo de <i>reads</i> permite a identificação dos dados de diferentes experimentos, ainda que eles estejam combinados em um arquivo único. Cada linha @RG deve conter um ID único. Logo, todas as <i>reads</i> de um mesmo grupo são consideradas como parte da mesma corrida de sequenciamento e indicam a linha que ocupam nessa corrida. Por exemplo: ID: FLOWCELL1. LANE2.
LB	identificador da biblioteca de DNA.
SM	nome da amostra sequenciada em determinado grupo de <i>reads</i> . Todo dado que tiver um mesmo valor SM será tratado como pertencendo a uma mesma amostra.

Continuação Tabela 2: Descrição dos componentes da linha de comando executada no BWA.

Comandos utilizados no BWA	Função do comando executado
PL	identifica a plataforma de sequenciamento utilizada.
PU	denominação alternativa do grupo de <i>reads</i> . Mantém a informação sobre a linha onde a amostra se encontra na <i>flow cell</i> . Essa designação não é exigida pelo GATK, mas prevalece sobre o ID caso seja executada uma recalibração de base.
/path/to/reference/hg19.fa	identifica o caminho a ser seguido para consulta ao genoma de referência
sample.trimmed.r1.fastq	<i>input</i> para o mapeamento, contendo a primeira <i>read</i> do par
sample.trimmed.r2.fastq	<i>input</i> para o mapeamento, contendo a segunda <i>read</i> do par
> mapped.sample.sam	indica o redirecionamento do resultado do mapeamento para o arquivo SAM especificado.

Como *output* do comando acima é gerado um arquivo SAM (*.sam). Porém, uma vez que os arquivos SAM são arquivos de texto que ocupam muito espaço no disco rígido, e para que as informações sobre mapeamento pudessem ser visualizadas, esse arquivo pode ser convertido para o formato BAM (*.bam).

1.2.3.1. SAM

O formato SAM (*Sequencing Alignment/Map*) (Li et al., 2009) é um arquivo baseado em texto, utilizado como *output* por diversos programas de alinhamento. Cada linha no arquivo SAM armazena informações sobre uma *read* mapeada contra um genoma de referência.

Arquivos SAM possuem uma seção com o cabeçalho (que é opcional) e uma seção de alinhamento. As linhas no cabeçalho iniciam com “@” seguido por um código de duas letras. Por sua vez, cada *read* alinhada à referência é representada em uma linha da seção de alinhamento.

Podem estar presentes ainda linhas opcionais com marcações predefinidas (*The SAM/BAM Format Specification Working Group, Sequence Alignment/Map Format*

Specification, 2015, <http://samtools.github.io/hts-specs/SAMv1.pdf>). Algumas destas marcações são geradas pelo BWA (aquelas começando com “X”).

1.2.3.2. BAM

O formato BAM (*Binary Alignment/Map*) é a versão binária compactada de um arquivo SAM, indexável e amplamente utilizado para representar as sequências após o alinhamento. A compactação em blocos (chamada BGZF) de um arquivo BAM permite além da diminuição do espaço de armazenamento o carregamento parcial do arquivo, o que é bastante utilizado em programas de visualização como o IGV ou o *UCSC Genome Browser*.

A conversão do arquivo SAM em BAM gera um arquivo com as mesmas informações e pode ser realizada através do programa *SAMtools* (<http://samtools.sourceforge.net/>), o qual fornece diversas ferramentas para manipulação desses arquivos.

Para visualizar o arquivo BAM e permitir sua utilização pelos *softwares* de determinação de variantes, é preciso ordenar os alinhamentos de acordo com a posição genômica. O comando a seguir é utilizado para realizar a conversão de SAM em BAM e a ordenação do arquivo BAM:

```
samtools sort mapped.sample.sam sorted.mapped.sample.bam
```

O nome do arquivo gerado pelo comando seguiu o formato a seguir, onde PIG0001 representa o nome da amostra e *sorted* indica que o arquivo BAM gerado se encontra ordenado:

```
PIG0001.sorted.bam
```

Arquivos BAM armazenam as mesmas informações que um arquivo SAM e possuem uma seção de cabeçalho, que contém informações gerais sobre o arquivo como nome, tamanho da amostra, método de alinhamento utilizado e uma seção com o alinhamento propriamente dito contendo nome, sequência e qualidade da *read*, além de marcações personalizadas.

Os arquivos BAM devem ainda ser indexados, permitindo uma pesquisa rápida no arquivo ordenado. Um índice de correspondência do arquivo BAM em questão em relação às coordenadas de referência será identificado pela terminação (*.bam.bai). O comando a seguir permitiu gerar um arquivo indexado:

```
samtools index sorted.mapped.sample.bam
```

O *output* do comando acima pode ser nomeado como a seguir:

```
FIG0001.sorted.bam.bai
```

Uma vez ordenados e indexados, os arquivos BAM puderam ser visualizados através de um visualizador externo. Neste trabalho utilizamos o *Integrative Genomics Viewer* (IGV) 2.3.67, um programa em java disponibilizado pelo *Broad Institute* (Robinson et al., 2011; Thorvaldsdottir, Robinson, & Mesirov, 2013).

1.3. Análise terciária

1.3.1. Determinação de variantes

Em um projeto de sequenciamento, o passo seguinte ao mapeamento é a detecção das diferenças entre as regiões genômicas sequenciadas e o genoma de referência (determinação de variantes). Um programa amplamente utilizado para análise dos dados de NGS é o *Genome Analysis Toolkit* (GATK) v.3.5, disponibilizado pelo *Broad Institute* (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013).

O programa apresenta uma estrutura modular e é dividido em diversas análises, as quais são reunidas junto a um *framework* em um arquivo java (*.jar). Para ter acesso à lista de análises disponíveis é utilizado o comando:

```
java -jar GenomeAnalysisTK.jar
```

Já para realizar uma análise específica é utilizado o comando:

```
java -jar GenomeAnalysisTK.jar -T <analysis> <arguments*>
```

Para realizar a determinação de SNPs e indels as ferramentas *HaplotypeCaller* com o parâmetro `--emitRefConfidence GVCF` e *UnifiedGenotyper* do GATK podem ser utilizadas. O *HaplotypeCaller* considera simultaneamente todas as amostras a serem analisadas e, ao encontrar uma região apresentando sinais de variação (região ativa), o programa descarta o mapeamento preexistente e realiza uma remontagem das *reads* naquela região. O programa realinha então cada sequência contra a sequência de referência com o objetivo de identificar potenciais sítios de variação.

Comparado ao *UnifiedGenotyper*, o *HaplotypeCaller* possibilita a nomeação de SNPs e indels de forma simultânea. Isso permite que o *HaplotypeCaller* seja mais preciso quando lidando com regiões reconhecidamente problemáticas, como por exemplo, onde são encontrados diferentes tipos de variações próximas umas das outras (Van der Auwera et al., 2013).

Por outro lado, o *UnifiedGenotyper* possui filtros menos restritivos que possibilitam encontrar variantes com frequências populacionais mais baixas, principalmente em áreas de baixa cobertura, como *singletons* e alelos raros em polimorfismos multialélicos. Porém, essa característica pode resultar em uma taxa alta de falsos positivos. Além disso, essa ferramenta faz a leitura das sequências ponto a ponto, não sendo aplicável à identificação de indels. Por essas e outras razões, o próprio site do GATK descontinuou as atualizações e aposentou a ferramenta *UnifiedGenotyper* recomendando o uso do *HaplotypeCaller* em seu lugar.

Devido a utilização do parâmetro `--emitRefConfidence GVCF`, é gerado para cada amostra separadamente um arquivo gVCF (do inglês, *genomic variant calling format*), onde estão dispostas um registro para cada posição (ou intervalo de interesse) independentemente de uma variante ser detectada naquele sítio ou não. O uso do gVCF é importante pois permite a adição de outros conjuntos de amostra, caso necessite realizar uma análise conjunta. Desta forma, pode-se adicionar os arquivos gVCFs de diferentes coortes e realizar a análise de chamada de variantes simultaneamente nestes conjuntos de amostra, aumentando a acurácia nesta determinação.

As seguintes linhas de comando são utilizadas para realizar a nominação das variantes (Tabela 3):

HaplotypeCaller:

```
java -jar GenomeAnalysisTK.jar \
  -R reference.fasta \
  -T HaplotypeCaller \
  -I listofsamples.list \
  --emitRefConfidence GVCF \
  --dbsnp dbSNP.vcf \
  -L targets.interval_list \
  -o output.raw.snps.indels.g.vcf \
  --dontUseSoftClippedBases \
  -drf DuplicateRead
```

UnifiedGenotyper:

```
java -jar GenomeAnalysisTK.jar \
  -T UnifiedGenotyper \
  -R reference.fasta \
  -I sample1.bam [-I sample2.bam ...] \
  --dbsnp dbSNP.vcf \
  -o snps.raw.vcf \
  -stand_call_conf [50.0] \
  -stand_emit_conf 10.0 \
  [-L targets.interval_list]
```

Tabela 3: Descrição dos componentes da linha de comando executada no GATK

Comandos utilizados no GATK	Função do comando executado
-R	indica o arquivo com o genoma de referência
-T	indica qual ferramenta do GATK será utilizada, isto é, <i>HaplotypeCaller</i> ou <i>UnifiedGenotyper</i>
-I	indica arquivo de texto com lista de amostras (formato BAM) a ser utilizado pelo GATK
--dbsnp	indica arquivo (dbSNP.vcf) contendo a identificação dos rs de cada ponto de variação oficialmente descrito, a ser utilizada para preenchimento da coluna ID do <i>output</i>
-stand_call_conf	é o limiar mínimo de confiança (na escala Phred) utilizado para separar determinações de genótipo realizadas com maior e menor confiança. Apenas determinações de genótipo emitidas com confiança $\geq$ limiar mínimo são consideradas. O valor default recomendado é de 30

Continuação Tabela 3: Descrição dos componentes da linha de comando executada no GATK

Comandos utilizados no GATK	Função do comando executado
-stand_emit_conf	permite emitir determinações de genótipo com baixa qualidade como resultados filtrados
--emitRefConfidence GVCF	indica a produção do arquivo g.vcf para o <i>HaplotypeCaller</i>
-L	direciona o GATK a restringir o processamento dos dados aos intervalos genômicos especificados em um arquivo “ <i>targets.interval_list</i> ”, que indica o cromossomo e as coordenadas das posições de início e fim em relação ao genoma de referência
-o	indica o <i>output</i> , que é um arquivo VCF ou gVCF (<i>Variant Call Format</i>) contendo os sítios que o algoritmo (<i>UnifiedGenotyper</i> ou <i>HaplotypeCaller</i> , respectivamente) identificou como sendo variáveis
--dontUseSoftClippedBases	direciona o programa a não analisar bases <i>soft clipped</i> (bases não alinhadas nas extremidades das <i>reads</i>), o que diminui a determinação de falso-positivos
-drf	desabilita a aplicação de filtros à análise (<i>DuplicateRead</i> neste caso).

Após gerar os arquivos g.vcf para cada amostra um passo é realizado para combinar os arquivos em um único arquivo vcf com a linha de comando:

```
java -jar GenomeAnalysisTK.jar \
  -T GenotypeGVCFs \
  -R reference.fasta \
  -I listofsamples.list \
  -o cohort.g.vcf
```

Tabela 4: Descrição dos componentes da linha de comando executada no GATK

Comandos utilizados no GATK	Função do comando executado
-R	indica o arquivo com o genoma de referência
-T	indica qual ferramenta do GATK será utilizada, isto é, <i>GenotypeGVCFs</i>
-I	indica arquivo de texto com lista de amostras (formato BAM) a ser utilizado pelo GATK
-o	indica o <i>output</i> , que é um arquivo gVCF (<i>Variant Call Format</i>) contendo os sítios que o algoritmo (<i>HaplotypeCaller</i>) identificou como sendo variáveis

1.3.1.1. VCF

O VCF é um formato de arquivo que contém informações sobre variantes encontradas em posições específicas em relação a um genoma de referência. O arquivo VCF possui um cabeçalho, bastante útil, uma vez que contém uma descrição para os campos contidos no arquivo (Tabela 5):

Tabela 5: Descrição dos componentes da seção de cabeçalho de um arquivo VCF

Informações presentes no cabeçalho	Descrição das abreviações apresentadas
Fileformat	versão do formato de arquivo e a versão do <i>variant caller</i> utilizado
FILTER	identifica quais filtros são aplicados aos dados
FORMAT e INFO	essas linhas definem as anotações contidas nas colunas correspondentes a essas informações no arquivo VCF
GATKCommandLine	contém a linha de comando utilizada pelo GATK para realizar a genotipagem das amostras. Essa linha de comando especifica todos os parâmetros utilizados pelo <i>variant caller</i> , incluindo os nomes do arquivo com o genoma de referência e do arquivo BAM
Contig lines e Reference	contém informações como os nomes dos fragmentos, tamanho e genoma de referência que é utilizado como <i>input</i> para gerar o arquivo BAM

Além do cabeçalho, um arquivo VCF contém as linhas com os dados das amostras, onde estão contidas em cada linha informações sobre uma única variante. As linhas com os dados são listadas seguindo o designado no cabeçalho do arquivo VCF. As oito primeiras colunas do arquivo (até INFO) representam as propriedades observadas dos sítios variantes (ou invariantes). Informações específicas à amostra são apresentadas na coluna FORMAT e nas seguintes.

As nove primeiras colunas são necessárias para o formato VCF, ainda que possam estar vazias. As informações contidas em cada coluna estão apresentadas na Tabela 6.

Tabela 6: Descrição dos componentes do corpo de um arquivo VCF

Informações presentes no corpo do arquivo	Descrição das abreviações apresentadas
CHROM e POS	contém as coordenadas genômicas nas quais

	as variantes ocorrem
ID	campo opcional, cujo conteúdo se baseia nas informações contidas em um banco de dados de referência como o dbSNP
REF e ALT	identificam o alelo referência e alternativo(s) observado(s) na(s) amostra(s). No caso de inserções, o alelo ALT fornece a sequência inserida, bem como a base anterior à inserção. Para deleções, o alelo ALT é representado pela base anterior à deleção
QUAL	é a probabilidade de que um polimorfismo (REF/ALT) exista de fato neste local indicado. No entanto, por ser este valor dependente da quantidade de dados analisada ele não representa uma medida útil para avaliação da qualidade de determinação de uma variante
FILTER	esse campo contém os filtros aplicados aos dados analisados. Se a variante passou pelos filtros aplicados o valor indicado é PASS. Caso nenhum filtro tenha sido aplicado será indicado “.”
INFO	informações descrevendo cada variante
DB	associado ao dbSNP
H3	associado ao HapMap3
VALIDATED	validado por experimentos contínuos
NA	número total de alelos nos genótipos identificados
AC	contagem alélica dentre os genótipos do alelo ALT, na mesma ordem em que listados
SVTYPE	tipo de variante estrutural (DEL para deleção, DUP para duplicação, INV para inversão)
END	posição final ocupada pela variante
IMPRECISE	indica que a posição ocupada pela variante não é precisa
CIPOS/CIEND	intervalo de confiança em torno da posição e posição final para variantes localizadas com baixa precisão
FORMAT	fornece a ordem das informações contidas nas colunas seguintes (amostras):
GT	indica o genótipo da amostra em cada sítio. Em diploides, são indicados os dois alelos de cada indivíduo, sendo “0” indicativo do alelo referência e “1” do alelo alternativo. O separador (/) indica que os genótipos não estão faseados e () indica genótipos em fase
AD	representa o número de <i>reads</i> que dá suporte a cada um dos alelos (alternativo e/ou referência) reportados
DP	indica o número total de <i>reads</i> que representam o genótipo reportado
GQ	indica qual a confiança (escala Phred) da determinação de genótipo realizada. Valores muito baixos são indicativos de pouca

	confiança no genótipo identificado
PL	é a probabilidade normalizada dos genótipos possíveis ou a probabilidade de o genótipo não estar correto. Assim, quanto menor o valor apresentado aqui, melhor
HP	identificadores de haplótipos eventualmente determinados pelo <i>ReadBackedPhasing</i>
PQ	qualidade de faseamento de haplótipos eventualmente determinados pelo <i>ReadBackedPhasing</i>

1.3.2. Remoção de genótipos de baixa qualidade

Alguns dos genótipos considerados pelo *HaplotypeCaller* e *UnifiedGenotyper* podem não ser verdadeiros, principalmente quando a genotipagem se refere a um segmento com baixa cobertura de sequenciamento ou quando, no caso de genótipos heterozigotos, um dos alelos é pouco representado.

Visando diminuir a taxa de falso-positivos, genótipos de baixa cobertura são interrogados, tomando como *input* o arquivo VCF obtido pelo *HaplotypeCaller* e *UnifiedGenotyper*.

Neste processo, o arquivo VCF é tratado pela ferramenta VCFx (versão 1.0) (Castelli et al., 2015), disponível em http://www.castelli-lab.net/apps/apps_vcfx.php, que interroga alelos segundo as regras:

- Para genótipos homozigotos inferidos em segmentos cobertos por oito ou menos *reads*, é introduzido um alelo interrogado (*vcfx alpha* = 8). A definição final desse tipo de genótipo (homozigoto ou heterozigoto) será inferida posteriormente.
- Genótipos heterozigotos em que um dos alelos está altamente subrepresentado (proporção de *reads* igual ou menor que 1%), são considerados homozigotos para o alelo mais representado (*vcfx beta* = 0.01). Esse procedimento minimiza a influência de *reads* mapeadas incorretamente e da alta taxa de erros de sequenciamento geralmente introduzidos em dados de NGS. Essa correção só pode ser aplicada em situações de alta cobertura (100 ou mais *reads*).
- Para genótipos heterozigotos em que um dos alelos está relativamente subrepresentado (proporção de *reads* entre 1% e 20%), um alelo interrogado é introduzido representando o alelo subrepresentado (*vcfx delta* = 0.20). A definição final desse tipo de genótipo (homozigoto ou heterozigoto) será inferida posteriormente.

- Para genótipos heterozigotos inferidos com no máximo cinco *reads* em que um dos alelos está subrepresentado (Proporção de *reads* ente 20% e 40%), um alelo interrogado é inserido representando o alelo subrepresentado. No entanto, quando os dois alelos têm a mesma representatividade, ambos são considerados interrogados (*vcfx gamma* = 0.50). A definição final desse tipo de genótipo (homozigoto ou heterozigoto) será inferida posteriormente.

A linha de comando utilizada no VCFx é indicada a seguir:

```
vcfx checkpl input=input.vcf output=output.vcf
```

Após a análise pelo VCFx, se forem realizadas análises por ambas as ferramentas (*HaplotypeCaller* e *UnifiedGenotyper*) as variantes que apresentaram genótipos distintos nos dois arquivos VCF obtidos através do *HaplotypeCaller* e *UnifiedGenotyper*, são analisadas manualmente para (a) verificar qual dos dois algoritmos fez a identificação correta do genótipo e (b) nos casos de genótipos interrogados por somente um dos algoritmos, para resolver variantes que são interrogadas por conta do número de *reads* consideradas pelos algoritmos, principalmente no caso do *HaplotypeCaller* que, devido aos filtros extremamente rigorosos, acaba descartando um número considerável de *reads* de suas análises. Para isso, é preciso analisar os arquivos BAM das amostras nos pontos interrogados. Como descrito anteriormente, o programa IGV pode ser utilizado para a visualização, avaliação de qualidade e contagem das *reads* alternativas que cobrem determinada posição. Considerando somente as bases chamadas com qualidade mínima Q30 em *reads* mapeadas com qualidade mínima Q30, as regras empregadas pelo VCFx são novamente aplicadas, agora manualmente para determinar se essas variantes continuariam interrogadas ou se seriam resolvidas.

Após essa etapa de maturação do arquivo, as informações contidas nos dois arquivos são unidas em um único arquivo VCF (VCF consenso) que é utilizado nas análises posteriores.

Após a análise pelo VCFx, se apenas a ferramenta *HaplotypeCaller* com o parâmetro -emitRefConfidence GVCF foi utilizada, os pontos de variação interrogados das amostras devem ser visualizados no programa IGV. É preciso analisar os arquivos BAMs destas amostras para avaliar a qualidade e contagem de *reads* alternativas que cobrem determinada posição (interrogada pelo programa VCFx). Isto porque o *HaplotypeCaller* aplica filtros

extremamente rigorosos, descartando um número considerável de *reads* para suas análises. Portanto, analisar ponto a ponto os arquivos BAM muitas vezes pode auxiliar na compreensão do porquê as variantes foram interrogadas pelo programa VCFx.

Após a maturação manual o arquivo VCF é atualizado manualmente nas posições onde houveram alelos interrogados reavaliados e resolvidos.

2. REFERÊNCIAS

- Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Castelli, E. C., Mendes-Junior, C. T., Sabbagh, A., Porto, I. O. P., Garcia, A., Ramalho, J., ... Donadi, E. A. (2015). HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Human Immunology*, 76(12), 945–953. <https://doi.org/10.1016/j.humimm.2015.06.016>
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–8. <https://doi.org/10.1038/ng.806>
- Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8(3), 175–185. <https://doi.org/10.1101/gr.8.3.175>
- Ferragina, P., Manzini, G., Veli, M., & Navarro, G. (2004). An Alphabet-Friendly FM-Index, 150–160.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1), 10–12. <https://doi.org/10.14806/ej.17.1.200>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>

- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192. <https://doi.org/10.1093/bib/bbs017>
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 11, p. 11.10.1-11.10.33). Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/0471250953.bi1110s43>