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

“Análise da acurácia de diferentes sistemas genéticos de predição de fenótipos de pigmentação em amostra de indivíduos miscigenados da população brasileira”.

Thássia Mayra Telles Carratto

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

2019

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Thássia Mayra Telles Carratto

Análise da acurácia de diferentes sistemas genéticos de predição de fenótipos de pigmentação
em amostra de indivíduos miscigenados da população brasileira

Versão Corrigida

Dissertação apresentada à Faculdade de Filosofia,
Ciências e Letras de Ribeirão Preto da
Universidade de São Paulo, como parte das
exigências para obtenção do título de Mestre em
Ciências, obtido no Programa de Pós-Graduação
em **Química**

Orientador: Prof. Dr. Celso Teixeira Mendes Junior

Ribeirão Preto - SP
2019

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.

Carratto, Thássia Mayra Telles

Análise da acurácia de diferentes sistemas genéticos de predição de fenótipos de pigmentação em amostra de indivíduos miscigenados da população brasileira, 2019.

144 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. Fenotipagem Forense. 2. Pigmentação Humana. 3. Ancestralidade.

Nome: Carratto, Thássia Mayra Telles

Título: Análise da acurácia de diferentes sistemas genéticos de predição de fenótipos de pigmentação em amostra de indivíduos miscigenados da população brasileira

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:

Banca Examinadora

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____

Julgamento: _____ Assinatura: _____

Para meus pais e meu avô, os maiores amores da minha vida

Agradecimentos

Primeiramente, gostaria de agradecer meus pais, Elaine e Fernando, por tudo que fizeram e fazem por mim até hoje. Por sempre me apoiarem em minhas escolhas, acreditarem em mim quando muitas vezes eu não acreditava e também por sempre terem batalhado para me proporcionar uma boa educação.

Agradeço também ao meu padrasto, José Luís e minha madrastra, Mônia, que sempre me tiveram como uma filha, cuidando de mim e me incentivando. Agradeço também ao meu irmão Gabriel que, apesar de criança, sempre me faz rir e me alegra com seu amor incondicional.

À toda a minha família, mas em especial aos meus avós, que são os amores da minha vida e estão sempre presentes, nunca medindo esforços para me ver feliz e me ajudar em tudo que for necessário.

Agradeço a todos os amigos que tive durante a graduação e o mestrado. Alguns já não fazem mais parte do meu dia-a-dia, outros chegaram recentemente, porém todos foram e ainda são de imensa importância para o meu desenvolvimento pessoal. Com certeza carregarei comigo lembranças muito boas de todos vocês. Gostaria de agradecer em especial minhas amigas Marcella, Natália, Emelin e Natalia, que sempre estão comigo, ainda que nem sempre fisicamente, prontas para conversar, aconselhar, dar risadas e me acalmar quando necessário. Não consigo imaginar uma vida sem vocês.

Aos meus clientes de trufa, que ao longo do tempo acabaram se tornando também meus amigos. Muito obrigada pelas conversas compartilhadas toda semana.

Agradeço também aos meus amigos do Laboratório de Pesquisas Forenses e Genômica: Letícia, Maria Luiza, Nádia, Guizão, Guilherme, Tamara, Amanda e Alison. Vocês são muito especiais para mim e agradeço imensamente toda a ajuda recebida, além das horas de conversa jogadas fora. Nunca imaginei ter meu grupo de laboratório como uma segunda família, e a isso sou muito grata. Agradeço especialmente a Letícia, que desde o meu primeiro dia no laboratório sempre me ajudou de prontidão e com muita paciência.

Agradeço meu orientador, prof. Dr. Celso, pela oportunidade de desenvolver uma pesquisa tão linda e apaixonante. Gostaria de expressar aqui a minha imensa admiração por ele, que é sempre justo e honesto. Me inspiro muito em sua pessoa e sou muito grata por toda a ajuda e paciência comigo, além das muitas conversas durante meus dias no laboratório.

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

Ao Prof. Dr. Érick da Cruz Castelli pelo suporte em bioinformática.

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.

Ao apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Aos membros da banca examinadora por aceitarem participar de minha defesa e compartilhar seus conhecimentos, contribuindo para este trabalho e para meu desenvolvimento profissional.

À todos os voluntários que participaram da pesquisa, obrigada pela sua disposição e paciência. Nosso trabalho só existe por causa de vocês.

Resumo

CARRATTO, T. M. T. **Análise da acurácia de diferentes sistemas genéticos de predição de fenótipos de pigmentação em amostra de indivíduos miscigenados da população brasileira.** 2019. 144 páginas. Dissertação de Mestrado – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil.

Buscando auxiliar a polícia em casos nos quais o perfil genético obtido tradicionalmente por STRs não é suficiente para solucionar um crime, foi proposta uma técnica, a qual consiste em determinar características externamente visíveis (EVCs) de um indivíduo baseando-se somente na análise de sítios de variação presentes em seu material genético, denominada *Forensic DNA Phenotyping* (FDP). Os fenótipos de pigmentação (cor de olho, pele e cabelo) se tornaram fortes candidatos a serem as primeiras EVCs a ter uma metodologia de predição testada e validada, pois são características de fácil identificação visual e atualmente já se tem uma boa compreensão sobre a base genética da pigmentação humana. Em 2011, foi desenvolvido o sistema *IrisPlex*, a primeira ferramenta testada e validada para predição da cor dos olhos, composto por seis SNPs. Posteriormente, 18 novos marcadores para predição da cor dos cabelos foram incluídos nesta ferramenta, dando origem ao sistema *HIrisPlex*. Por fim, em 2017 foi desenvolvido o sistema *HIrisPlex-S*, uma ferramenta composta por 41 marcadores capaz de prever simultaneamente a cor dos olhos, cabelos e pele de um indivíduo. Paralelamente, foram desenvolvidos três modelos para predição da cor dos olhos (compostos por 7, 13 e 23 marcadores), um modelo para predição da cor dos cabelos (composto por 12 SNPs) e um para predição da cor da pele (composto por 10 SNPs), todos vinculados a uma ferramenta denominada *Snipper app suite version 2.5*. Como as ferramentas preditivas mencionadas aqui foram desenvolvidas utilizando como base populações europeias, temos por hipótese que esses sistemas preditivos não terão êxito ao serem testados em populações altamente miscigenadas, como a brasileira, falhando principalmente na predição de fenótipos intermediários. Deste modo, o presente projeto tem como objetivo testar os conjuntos de predição fenotípica em uma amostra de 501 indivíduos da população brasileira, avaliando suas acurácias específica e geral, a associação individual dos marcadores com as características morfológicas e verificar a utilidade dos conjuntos na inferência de ancestralidade, outro elemento de grande interesse em investigações. Os genótipos dos 61 marcadores que compõem todas as ferramentas mencionadas acima foram obtidos por sequenciamento de nova geração (NGS), genotipagem em larga escala (MEGA Array) e imputação utilizando os programas

IMPUTE2 e *Michigan Imputation Server*. As proporções de ancestralidade foram inferidas utilizando o *software* STRUCTURE. As análises preditivas foram realizadas por meio das ferramentas online *HIrisPlex-S* e *Snipper*, enquanto as associações foram identificadas por meio de uma regressão logística realizada pelo *software* plink v1.07, adicionando ou não a ancestralidade de cada indivíduo como fator de confusão. Os conjuntos de marcadores não se mostraram úteis na inferência da ancestralidade, com exceção do *Snipper10*, e apenas 38 marcadores apresentaram alguma associação com os fenótipos de pigmentação. Em relação à cor dos olhos e cabelos, o sistema *HIrisPlex-S* apresentou sensibilidade geral maior que os conjuntos *Snipper*, porém ambas as ferramentas apresentam falhas na predição de fenótipos intermediários. Em relação à cor da pele, o sistema *HIrisPlex-S* apresentou melhores resultados na predição de peles muito escuras, enquanto o *Snipper10* apresentou bons resultados para peles brancas. Apesar disso, nenhuma das ferramentas apresentou performance boa o suficiente para serem usadas com confiabilidade na população brasileira. Este trabalho ressalta que a busca de marcadores associados com fenótipos de pigmentação em populações miscigenadas é de extrema importância para que se possa criar uma ferramenta de maior valor preditivo e que possa ser usada com confiabilidade em tais populações.

Palavras-chave: Fenotipagem forense; Pigmentação humana; Ancestralidade.

Abstract

CARRATTO, T. M. T. **Evaluation of accuracy of different genetics systems available for prediction of pigmentation phenotypes in an admixed Brazilian population sample.** 2019. 144 pages. Dissertation (Master's Degree) - Faculty of Philosophy, Sciences and Letters at Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brasil.

Aiming to help the police in cases in which a standard genetic profile obtained using STRs cannot solve a crime, a technic called Forensic DNA Phenotyping (FDP) was proposed. FDP consists in determining externally visible characteristics (EVCs) of an individual from a set of DNA markers. The pigmentation phenotypes (eye, hair and skin color) became strong candidates to be the first EVCs with a prediction methodology tested and validated, because these characteristics are of easy visual identification, in addition to the fact that nowadays we have a proper comprehension of the human pigmentation genetic basis. In 2011, the IrisPlex system was developed. This tool was the first one to be tested and validated, and it was designed to predict eye color using six SNPs. Later, 18 markers were added in the system in order to predict hair color, composing the HIrisPlex system. Finally, in 2017, the HIrisPlex-S system was developed. This tool is composed of 41 markers and it aims to predict eye, hair and skin color simultaneously. Along with that, three other sets of markers for eye color prediction (composed of 7, 13 and 23 SNPs) were developed, in addition to one set for hair (composed of 12 SNPs) and one set for skin color prediction (composed of 10 SNPs). All these tools are hosted in the Snipper app suite version 2.5. As the predictive systems named here were developed based in studies involving European populations, we have the hypothesis that these tools will not present a satisfactory performance in highly admixed populations, as the Brazilian one, failing mainly in predicting intermediate phenotypes. In this way, the actual project aims to test the phenotypes prediction systems in a 501 Brazilian admixed population sample, evaluating their predictive accuracies, as well as the association of specific markers with morphological characteristics. In addition, we also aim to verify if the tools can be useful to infer ancestry, another element of great interest in investigations. The genotypes of the 61 markers that compose all the predictive tools were obtained by next-generation sequencing (NGS), genome-wide genotyping array (MEGA Array) and imputation procedures using the IMPUTE2 and Michigan Imputation Server programs. The software STRUCTURE was used to infer ancestry proportions. Predictive analyses were performed using the HIrisPlex-S and Snipper online tools. A logistic regression model performed by the software plink v1.07 identified associations between markers and pigmentation characteristics, adjusting or not for individual genomic ancestry. The predictive tools were not helpful in inferring ancestry, except for Snipper10. Only 38 out of 61 markers showed any association with pigmentation phenotypes. Regarding eye and hair color, the HIrisPlex-S system presented higher overall sensitivity than the Snipper sets, but all of them failed in predicting intermediate phenotypes. Regarding skin color, the HIrisPlex-S system presented better results in predicting dark-to-black skin, while Snipper10 presented good results for white skin. Despite that, none of the

tools displayed a performance good enough to support their use with reliability in the admixed Brazilian population. This study highlights the extreme importance of searching new markers that are associated with pigmentation phenotypes in admixed populations, in order to develop a tool with higher predictive value that will be able to be used in such populations.

Keywords: Forensic phenotyping; Human pigmentation; Ancestry.

Lista de Figuras

Figura 1 – Processo da síntese da eumelanina e feomelanina que ocorre nos melanócitos. Figura criada com o auxílio da ferramenta ACD/ChemSketch (ACD/Labs Inc.) e retirada de Marcorin 2017.	18
Figura 2 – Exemplo da classificação da coloração de um objeto utilizando o espaço de cor CIE L*a*b*. Retirado de http://sensing.konicaminolta.com.br/2013/11/entendendo-o-espaco-de-cor-lab/	21
Figura 3 – Três curvas ROC de diferentes classificadores preditivos. A curva verde corresponde à uma ferramenta com predição perfeita, enquanto a curva azul corresponde à uma ferramenta com predição boa e a curva vermelha à uma ferramenta com poder de predição aleatório. Retirado de http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/	45
Figura 4 – Quadro delineando a distribuição dos resultados gerados por uma ferramenta preditiva na análise de um fenótipo binário.	46
Figura 5 – Curva ROC construída para a cor de cor de cabelo ruivo utilizando o sistema <i>HirisPlex-S</i> (AUC = 0,985).	60
Figura 6 – Curva ROC construída para a cor de olho intermediária utilizando o sistema <i>Snipper7</i> , MLR como classificadora e sem considerar haplótipo (AUC = 0,373).....	63
Figura 7 – Diagramas triangulares gerados pelo <i>software</i> STRUCTURE a partir dos seis sistemas de marcadores preditivos analisados, representando as estimativas de ancestralidade geradas para cada indivíduo. Os círculos coloridos representam os indivíduos, divididos em populações por meio das cores: amarelo – indivíduos provenientes de Ribeirão Preto, vermelho – europeus (EUR), azul – asiáticos (EAS), verde – africanos (AFR).....	73
Figura 8 – Diagrama triangular gerado pelo <i>software</i> STRUCTURE a partir dos 61 marcadores preditivos analisados, representando as estimativas de ancestralidade geradas para cada indivíduo. ..	74
Figura 9 – Gráficos retangulares gerados pelo <i>software</i> distruct 1.1, baseados nas análises realizadas pelo <i>software</i> STRUCTURE a partir dos seis sistemas de marcadores preditivos analisados, representando as estimativas de ancestralidade geradas para cada indivíduo. As barras verticais representam cada indivíduo e assumem até três cores, de acordo com a ancestralidade do mesmo: vermelho, verde e azul correspondem a ancestralidade europeia (EUR), africana (AFR) e asiática (EAS), respectivamente.	74

Lista de Tabelas

Tabela 1 – Regiões imputadas pelos <i>softwares</i> IMPUTE2 e <i>Michigan Imputation Server</i>	41
Tabela 2 – SNPs estudados, incluindo informações referentes à localização cromossômica e gênica, meio de obtenção dos genótipos de cada indivíduo e distribuição em painéis de SNPs para predição de fenótipos de pigmentação.....	42
Tabela 3 – Marcadores escolhidos para validação dos resultados, qualidade teórica da imputação fornecida pelos <i>softwares</i> e acurácia real da imputação mensurada pela taxa de concordância com os genótipos determinados por NGS.	52
Tabela 4 – Marcadores imputados e a qualidade da imputação fornecida por cada ferramenta.	53
Tabela 5 – Discordâncias encontradas nos resultados obtidos pelos dois métodos de imputação.	54
Tabela 6 – Número de genótipos não definidos pelo IMPUTE2, de discordâncias entre as duas ferramentas de imputação e de amostras com genótipos concordantes	54
Tabela 7 – Frequências alélicas, heterozigoses observada e esperada (HO e HE) e aderência ao Equilíbrio de Hardy-Weinberg (pEHW) referentes aos 61 marcadores de estudo. Desvios do equilíbrio de Hardy-Weinberg estão destacados em negrito. Os marcadores estão ordenados pela posição cromossômica que ocupam.	55
Tabela 8 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas existentes no sistema <i>HIrisPlex-S</i>	58
Tabela 9 – Sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema <i>HIrisPlex-S</i> , aplicando <i>threshold</i> de 0,5 e 0,7....	58
Tabela 10 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do conjunto <i>Snipper6</i> , considerando ou não o <i>cut-off</i> 3:1....	61
Tabela 11 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema <i>Snipper7</i> , considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o <i>cut-off</i> 3:1.	62
Tabela 12 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema <i>Snipper13</i> , considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o <i>cut-off</i> 3:1.	64
Tabela 13 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema <i>Snipper23</i> , considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o <i>cut-off</i> 3:1.	65
Tabela 14 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos cabelos, a partir do sistema <i>Snipper12</i> , considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o <i>cut-off</i> 3:1.	67
Tabela 15 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor de pele, a partir do sistema <i>Snipper10</i> , considerando ou não o <i>cut-off</i> 3:1.	68

Tabela 16 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor de pele, a partir do sistema <i>Snipper10</i> , considerando ou não o <i>cut-off</i> 3:1 e estratificando os indivíduos de acordo com a classificação de Fitzpatrick.	69
Tabela 17 – Associações alélicas encontradas para os 61 marcadores dos seis conjuntos preditivos analisados, adicionando a ancestralidade como um fator de confusão.	71
Tabela 18 – Estimativas de ancestralidade para a amostra populacional de Ribeirão Preto – SP, inferidas pelo <i>software</i> STRUCTURE, utilizando os marcadores presentes nas seis ferramentas preditivas analisadas.	75
Tabela 19 – AUCs estimados pelos desenvolvedores do sistema <i>HIrisPlex-S</i> e sensibilidades obtidas para os sistemas <i>Snipper</i> para as diferentes categorias de cada característica fenotípica em comparação com os valores obtidos no presente estudo.	80

Sumário

Sumário

1. Introdução	16
1.1 PIGMENTAÇÃO HUMANA.....	17
1.2 FENOTIPAGEM FORENSE POR DNA (<i>FORENSIC DNA PHENOTYPING</i> - FDP)	21
1.3 POPULAÇÃO BRASILEIRA	25
2. Justificativa	28
3. Hipótese	30
4. Objetivos	32
4.1 OBJETIVO GERAL	33
4.2 OBJETIVOS ESPECÍFICOS	33
5. Material e Métodos	34
5.1 COLETA	35
5.2 PROCEDIMENTOS LABORATORIAIS	36
5.2.1 Extração (Anexo C – Item 1)	36
5.2.2 Quantificação (Anexo C – Item 2)	36
5.2.3 Preparo e validação das bibliotecas (Anexo C – Itens 3 a 6)	36
5.2.4 Sequenciamento de nova geração (NGS) (Anexo C – Item 7).....	37
5.2.5 Genotipagem em larga escala por Array	37
5.3 PROCEDIMENTOS COMPUTACIONAIS (Anexo D)	38
5.3.1 Trimagem e mapeamento dos dados obtidos por NGS	38
5.3.2 Obtenção de genótipos	39
5.3.3 Análises populacionais.....	43
5.3.4 Análises de predições fenotípicas	44
5.3.4.1 <i>HIrisPlex-S</i>	46
5.3.4.2 <i>Snipper</i>	488
5.3.5 Análises de Ancestralidade	49
6. Resultados	50
6.1 GENOTIPAGEM.....	51
6.2 IMPUTAÇÃO	52
6.3 ANÁLISES POPULACIONAIS	54
6.4 ANÁLISES FENOTÍPICAS	57
6.4.1 <i>HIrisPlex-S</i>	57
6.4.2 <i>Snipper6</i> , <i>Snipper7</i> , <i>Snipper13</i> e <i>Snipper23</i>	60
6.4.3 <i>Snipper12</i>	66
6.4.4 <i>Snipper10</i>	67
6.5 ANÁLISES DE ASSOCIAÇÃO.....	69

6.6 ANÁLISES DE ANCESTRALIDADE	72
7. Discussão.....	76
7.1 IMPUTAÇÃO	77
7.2 DESVIOS DO EQUILÍBRIO DE HARDY-WEINBERG.....	78
7.3 PREDIÇÕES FENOTÍPICAS.....	79
7.3.1 – Cor dos olhos.....	80
7.3.2 – Cor dos cabelos	83
7.3.3 – Cor da pele	84
7.3.4 – Cor da pele, olhos e cabelos.....	87
7.4 ASSOCIAÇÕES FENOTÍPICAS	888
7.5 ANCESTRALIDADE GENÔMICA	89
8. Conclusão	922
Referências.....	944
Apêndices.....	104
Apêndice A.....	1055
Apêndice B.....	1066
Apêndice C.....	1111
Anexos.....	114
Anexo A – Termo de consentimento livre e esclarecido.....	1155
Anexo B – Termo de consentimento livre e esclarecido para armazenamento de material biológico	1177
Anexo C – Protocolo de análise laboratorial (HaloPlex) para sequenciamento de nova geração	1199
Anexo D – Protocolo de processamento computacional de dados de sequenciamento de nova geração	1277

Introdução

1. Introdução

1.1 PIGMENTAÇÃO HUMANA

A pigmentação humana é um traço altamente variável, tendo assim grande utilidade na diferenciação de indivíduos (Parra 2007; Liu et al. 2013). Essa característica é definida majoritariamente pela presença de melanina na epiderme, cabelo e íris (Liu et al. 2013). Esse pigmento consiste em um complexo grupo de biopolímeros sintetizados dentro de células denominadas melanócitos, as quais estão localizadas na camada basal da epiderme, assim como no bulbo capilar e íris (Parra 2007).

O processo no qual o aminoácido tirosina é usado como substrato para a síntese de melanina é denominado melanogênese e ocorre nos melanossomos, organelas componentes dos melanócitos. Esse processo é composto por uma cadeia de complexas reações químicas, a qual pode seguir diferentes caminhos, originando assim diferentes formas de melanina. As formas mais comuns são a eumelanina, pigmento que origina os fenótipos de tons marrons/preto, e a feomelanina, pigmento que origina os fenótipos de tons vermelhos/amarelos, além de também ser responsável pela presença de sardas (Parra 2007; Liu et al. 2013). No caso da pele e cabelos, os melanossomos maduros repletos de melanina são transferidos dos melanócitos para os queratinócitos adjacentes. No caso dos olhos, os melanossomos permanecem dentro dos melanócitos (Sturm et al. 1998; Parra 2007).

A Figura 1 ilustra a biossíntese da melanina. Sabe-se que o processo de melanogênese pode ser regulado, sendo o receptor de melanocortina 1 (MC1R) o regulador positivo mais importante, que é estimulado pela ligação do hormônio alfa-estimulante de melanócito (α MSH). Ele possui a função de ativar a proteína de ligação responsiva ao cAMP (CREB), aumentando a expressão da proteína MITF, que, ao ser ativada por fosforilação (P), estimula a transcrição da tirosinase (TYR), proteína relacionada à tirosinase 1 (TYRP1) e dopacromo tautomerase (DCT). A tirosinase, por sua vez, é a enzima principal para o início da biossíntese da melanina, pois catalisa as duas primeiras reações da cadeia: hidroxilação da tirosina em 3, 4- dihidroxifenilalanina (DOPA) e sua subsequente oxidação em DOPAquinona. A partir daí a DOPAquinona pode seguir dois caminhos reacionais diferentes, o que é ditado pela disponibilidade de cisteína no meio. Dispondo de cisteína, a DOPAquinona é transformada em

cisteinil-DOPA, posteriormente oxidada e polimerizada, levando à formação de feomelanina. Já na ausência de cisteína, a DOPAquinona se transforma espontaneamente em DOPAcromo, que pode gerar dois compostos: DHICA (ácido-5,6-diidroxiindol-2-carboxílico), resultado de uma isomerização catalisada pelo DCT; e DHI (5,6-dihidroxiindol), resultado de uma oxidação espontânea do DOPAcromo. Ambos os compostos são oxidados posteriormente, levando à formação de DHICA-melanina, responsável pela coloração marrom da melanina, e DHI-melanina, responsável pela coloração preta da melanina (Parra 2007; Cichorek et al. 2013; Liu et al. 2013).

Figura 1 – Processo da síntese da eumelanina e feomelanina que ocorre nos melanócitos. Figura criada com o auxílio da ferramenta ACD/ChemSketch (ACD/Labs Inc.) e retirada de Marcorin 2017.

Apesar do número de melanócitos presentes nos tecidos humanos não variar, a regulação da melanogênese pode ser diferente entre os indivíduos, assim como nos tecidos, alterando a composição dos melanócitos em termos de quantidade, tipo e distribuição de melanina, o que explica a variação fenotípica apresentada entre duas pessoas, assim como diferentes pigmentações em um mesmo indivíduo (por exemplo, uma pessoa de cabelos pretos e pele muito clara) (Liu et al. 2013). No geral, peles mais pigmentadas contêm muitas partículas melanossomais únicas, grandes e em forma de elipse, preenchidas por eumelanina. Já a pigmentação clara está associada com melanossomos menores, redondos e menos densos que

se agrupam, preenchidos por feomelanina (Sturm et al. 2001). Outros fatores podem influenciar na expressão gênica, como marcações epigenéticas e o ambiente externo do indivíduo (Sturm et al. 2001).

A melanina desempenha, dentre outras funções (Hill 1992), um papel importante na proteção da pele contra raios UV (Parra 2007). Dois estudos, um realizado por Biasutti (1953) e outro por Relethford (1997), mostraram que a cor da pele e a latitude estão fortemente relacionadas, sugerindo que ao longo dos anos o ambiente pode ter levado a uma adaptação genética. A hipótese evolutiva mais aceita foi proposta por Loomis (1967), na qual a correlação observada entre cor de pele e latitude tem como causa a necessidade da síntese de vitamina D no corpo humano. Isso está diretamente ligado à absorção da radiação UV, elemento necessário para a produção da vitamina. Além disso, o ácido fólico, necessário na reparação e síntese do DNA, além de ter diversas outras funções, é muito sensível a radiação UV. Dessa forma, indivíduos de pele clara estariam em desvantagem em zonas de alta radiação UV, apresentando deficiência de ácido fólico. Assim, a melanina desempenha seu papel controlando a entrada dos raios UV na pele e, portanto, os níveis de vitamina D sintetizados e a concentração ótima de ácido fólico nas células. Em zonas equatoriais, a radiação UV é mais forte e uma maior quantidade de melanina para sua proteção é vantajosa, levando assim à manutenção de indivíduos de peles mais escuras nessas regiões, enquanto em regiões distantes das zonas equatoriais há uma radiação UV mais fraca, o que torna vantajoso uma menor produção de melanina e, conseqüentemente, indivíduos com peles mais claras apresentam maior sucesso reprodutivo nessas áreas (Parra 2007; Liu et al. 2013).

Em relação à cor dos cabelos e olhos, é mais difícil explicar a variação fenotípica por meio de teorias adaptativas. No caso dos cabelos, porque a pele protegida pelos fios é muito pequena, dificilmente levando a um favorecimento de fios mais escuros perto de zonas equatoriais e fios claros em regiões de maiores latitudes; no caso dos olhos, porque o fundo da íris é altamente pigmentado, independentemente da coloração da parte externa, protegendo assim os olhos contra a luz. Assim, a hipótese de seleção sexual e também de escolha preferencial de um companheiro (*mate choice preference*) são as mais aceitas (Liu et al. 2013). Porém, é importante ressaltar que existe a possibilidade de que os marcadores associados com a cor da pele também estejam associados à cor de olhos/cabelo e/ou em desequilíbrio de ligação com marcadores associados à essas características. Assim, as hipóteses citadas nesse parágrafo podem ter mínima contribuição para o surgimento de tais fenótipos.

Conforme a biossíntese da melanina e os processos evolucionários foram sendo melhor compreendidos, surgiu também um melhor entendimento sobre os genes que estão, ou ao menos possuem grande probabilidade de estar, envolvidos na melanogênese, influenciando a pigmentação externa de um indivíduo. Atualmente, estima-se que mais de 100 genes estão atuando na determinação da pigmentação, dentre eles *TYR*, *TYRP1*, *OCA2*, *SLC45A2*, *SLC24A5*, *MC1R*, *ASIP*, *KITLG*, *HERC2*, *SLC24A4*, *IRF4*, *TPCN2* e *BNC2* (Parra 2007; Sturm 2009; Liu et al. 2013). Assim, estudos buscando encontrar polimorfismos que poderiam estar associados à pigmentação foram iniciados em vários lugares do mundo.

Estudos envolvendo alguns desses genes já foram feitos na população brasileira. Enquanto que sítios de variação presentes nos genes *SLC24A5*, *SLC45A2*, *OCA2* e *HERC2* tem influência na pigmentação da população brasileira e podem ser considerados para a prática forense, o gene *ASIP* requer estudos adicionais para confirmar se a sua contribuição para a pigmentação ocorre apenas em populações autóctones ou se está presente também em populações miscigenadas, como é o caso da brasileira (de Araújo Lima et al. 2015; Andrade et al. 2017; Fracasso et al. 2017). Foi descoberto também um sítio de variação no gene *IRF4* que está relacionado ao fenômeno de agrisalamento precoce dos cabelos (Adhikari et al. 2016), achado este corroborado por estudo desenvolvido em nosso laboratório (Oliveira 2017).

Classificando os indivíduos de acordo com sua cor de olhos, cabelos e pele, podemos estudar a pigmentação humana como um traço qualitativo. Para tanto, a maioria dos estudos utiliza a classificação de Fitzpatrick (1988), a qual se baseia nas características da pele associadas à reação da mesma à exposição ao sol. Porém, é possível quantificar a pigmentação presente, estudando-a também como um traço quantitativo. Os primeiros estudos que buscavam padronizar a medida da cor de pele foram feitos utilizando técnicas de concordância de cores, comparando a cor da pele do indivíduo com uma escala pré-determinada variando do branco ao preto (Robins 1991). Porém, com a invenção dos espectrômetros de refletância portáteis nos anos 50, métodos mais objetivos foram desenvolvidos e colocados em uso. Atualmente existem várias maneiras de estudar a genética da pigmentação de maneira quantitativa, utilizando, por exemplo, espectroscopia de refletância difusa, HPLC, colorimetria tricromática e fotos digitais (Parra 2007). O modelo mais aceito nos dias de hoje é o sistema CIE $L^*a^*b^*$ (CIELAB), estabelecido em 1976 pela *Commission Internationale de l'Eclairage* (CIE). Esse sistema parte da premissa de que qualquer cor pode ser representada por meio de três variáveis, medidas por espectrofotometria: L^* representa o eixo claro-escuro; a^* o eixo vermelho-verde; b^* o eixo azul-amarelo. Essas três variáveis podem ser colocadas em um eixo tridimensional (Weatherall

e Coombs 1992). A Figura 2 utiliza uma maçã a título de exemplificação: com os valores de a^* e b^* , a cor correspondente à fruta pode ser encontrada num espaço de cores. O valor de L^* corresponderia ao eixo z (perpendicular ao plano bidimensional ilustrado na figura) e proporcionaria informações sobre a quantidade de luz captada no objeto.

Figura 2 – Exemplo da classificação da coloração de um objeto utilizando o espaço de cor CIE $L^*a^*b^*$. Retirado de <http://sensing.konicaminolta.com.br/2013/11/entendendo-o-espaco-de-cor-lab/>.

1.2 FENOTIPAGEM FORENSE POR DNA (*FORENSIC DNA PHENOTYPING - FDP*)

Em relação ao uso de DNA no âmbito forense, a obtenção do perfil genético de um indivíduo utilizando marcadores STRs (*Short Tandem Repeats*) é o método que vem sendo utilizado por quase 20 anos para identificação humana. O perfil de STRs obtido em um material biológico encontrado em uma cena de crime pode ser comparado com o perfil de algum suspeito já identificado pela polícia ou com perfis já existentes no banco de dados nacional. Em desastres de massa essa genotipagem também pode se mostrar útil, pois o perfil obtido de uma vítima ainda não identificada pode ser comparado com o perfil de parentes ou com amostras de referência antemortem das vítimas do desastre.

Apesar de seu uso disseminado, muitas vezes o DNA encontrado (seja em cenas de crime ou em desastres) está degradado e em quantidade reduzida, dificultando a obtenção de um perfil de STRs completo. Nesses casos, estratégias alternativas seriam necessárias. Assim, o uso de SNPs (*Single Nucleotide Polymorphisms*) se tornou uma possibilidade interessante devido a algumas vantagens apresentadas em relação aos STRs. Em caso de amostras degradadas, por exemplo, a amplificação dos fragmentos de DNA se torna mais fácil devido ao fato do SNP ser um polimorfismo de uma única base, resultando em amplicons de PCR de tamanhos menores, o que não é possível em uma amplificação de STRs, que são altamente polimórficos e podem assumir diversos números de repetição, gerando inevitavelmente amplicons maiores. Além disso, em uma análise com SNPs não são encontrados picos *stutters*, o que normalmente dificulta análises por STRs (Kayser e de Knijff 2011). Apesar de possuírem um menor poder de discriminação, os SNPs são muito abundantes no genoma, e com as atuais técnicas multiplex é possível genotipar milhares deles, fazendo com que o poder de discriminação combinado seja alto. Estima-se que para se obter o mesmo poder de discriminação ou exclusão dos kits de STRs atuais, é necessário um número aproximadamente quatro vezes maior de SNPs (Goodwin et al. 2011).

Em outras situações, ainda que um perfil de STRs de excelente qualidade seja obtido, pode ser que o indivíduo a quem esse perfil pertence nunca seja encontrado, quer seja porque a polícia não o teve como suspeito ou porque seu DNA não estava presente no banco nacional de perfis genéticos (Kayser e Schneider 2009; Kayser 2015). Tais situações estimularam o aparecimento de um novo campo de estudo na área forense denominado fenotipagem forense por DNA (FDP), o qual tem por objetivo determinar as características externamente visíveis (EVCs) de um indivíduo baseando-se apenas na análise de SNPs do material genético obtido (Kayser 2015). Assim, quando a análise por STRs se mostrasse inconclusiva, a FDP traria um novo rumo à investigação, concentrando a busca da polícia em indivíduos que apresentassem determinados fenótipos ou revelando alguns aspectos da aparência antemortem de vítimas encontradas em desastres de massa (Kayser e de Knijff 2011).

O desenvolvimento de metodologias para fenotipagem por DNA ainda avança com dificuldades. Isto porque a genética que origina fenótipos ainda não é totalmente esclarecida. Além disso, todas as EVCs são consideradas traços complexos, ou seja, fatores ambientais, em conjunto com diversos genes, atuam na determinação de um fenótipo, o que torna ainda mais complicado o seu estudo. Porém, a base molecular de algumas características é menos complicada do que de outras, e conforme foram realizados estudos de associação genômica

ampla (GWAS) buscando associações fenotípicas, se tornou possível começar a esboçar algumas ferramentas preditivas (Kayser 2015).

Atualmente, já se tem uma boa compreensão sobre a base genética da pigmentação humana, e a mesma se mostra como um dos traços fenotípicos menos complicados. Com base nessas informações, em 2011 foi desenvolvido o sistema *IrisPlex*, a primeira ferramenta testada e validada para predição da cor dos olhos (Walsh, Liu, et al. 2011; Walsh, Lindenberg, et al. 2011). Composta por seis SNPs altamente informativos em relação à coloração da íris humana, a ferramenta classifica os olhos em três categorias: azul, intermediário e castanho. Seus desenvolvedores alegam que a acurácia para determinação de olhos castanhos e azuis está acima de 90% (Walsh et al. 2012).

Ao sistema *IrisPlex* foram adicionados 18 novos marcadores (17 SNPs e 1 InDel), objetivando a predição simultânea da cor dos olhos e cabelos. Dos 24 marcadores do sistema que passou a ser designado *HIrisPlex*, 22 são utilizados para predição da cor dos cabelos, enquanto apenas os mesmos seis originalmente usados no *IrisPlex* continuam sendo utilizados para predição da cor dos olhos (quatro são usados nas duas predições) (Walsh et al. 2013; Walsh et al. 2014). Os criadores do sistema desenvolveram um guia para a interpretação das probabilidades obtidas para cor do cabelo, criando assim cinco categorias: ruivo, loiro-claro, loiro-escuro/castanho-claro, castanho-escuro e preto (Walsh et al. 2014). Esse novo sistema foi testado e validado na Europa (Irlanda, Polônia e Grécia), apresentando valores de sensibilidade de 69,5% para cabelos loiros, 78,5% para cabelos castanhos, 80,0% para cabelos ruivos e 87,5% para cabelos pretos, com sensibilidade geral de 73,0% para a predição da cor de cabelo. Em relação aos olhos, a sensibilidade obtida foi de 84,0%, pois a dificuldade da ferramenta na predição de olhos intermediários leva à uma redução desse valor (Walsh, Liu, et al. 2011; Walsh, Lindenberg, et al. 2011). Dada a boa performance da ferramenta, mais marcadores foram adicionados, dando origem ao sistema *HIrisPlex-S*, composto por 41 marcadores, 36 deles sendo utilizados para predição da cor da pele (Chaitanya et al. 2018). Para estratificação da cor da pele, o sistema se baseia na classificação de Fitzpatrick (tipos I a VI) (1988), possuindo cinco categorias: muito clara (tipo I), clara (tipo II), intermediária (tipos III e IV), escura (tipo V) e muito escura (tipo VI). A ferramenta ainda está sendo testada e validada por seus criadores, porém os primeiros estudos em diferentes populações europeias demonstraram uma sensibilidade de 99,0% para pele clara (tipos I a IV), 50,0% para pele morena (tipo V) e 92,0% para pele escura (tipo VI) (Walsh et al. 2017).

Paralelamente, um grupo diferente de pesquisadores também desenvolveu outros conjuntos de sistemas preditivos vinculados a uma ferramenta denominada *Snipper app suite version 2.5* (<http://mathgene.usc.es/Snipper/>), que traz como principal inovação a utilização de um algoritmo Bayesiano de classificação. Dentre os sistemas, três modelos foram desenvolvidos para predição de cor dos olhos (Ruiz et al. 2013). O primeiro deles (aqui denominado *Snipper7*) realiza a predição de acordo com os genótipos de sete SNPs, seis dos quais estão presentes no sistema *IrisPlex*. O segundo modelo inclui outros seis marcadores, dando origem a um modelo baseado em 13 SNPs (*Snipper13*). Por fim, a esse segundo modelo foram adicionados 10 marcadores, gerando assim um terceiro modelo com 23 SNPs (*Snipper23*). No estudo original, estão disponíveis apenas os dados de sensibilidade atingidos pelo *Snipper23* quando testado por *cross-validation*: 97,93% para olhos azuis, 98,9% para intermediários e 92,1% para olhos castanhos. Para predição da cor de cabelo, foi desenvolvido um único modelo que utiliza genótipos de 12 SNPs (*Snipper12*), estando sete presentes no sistema *HIrisPlex* (Söchtig et al. 2015). As sensibilidades obtidas pelos desenvolvedores utilizando esse conjunto foram 77,8% para cabelos ruivos, 84,7% para loiros, 45,4% para castanhos e 75,0% para pretos. Por fim, para predição da cor da pele, 10 SNPs são utilizados, com quatro deles estando presentes também no *HIrisPlex-S* (Maroñas et al. 2014). Esse último modelo utiliza o valor de L* (sistema CIELAB) para classificar os indivíduos em cor de pele clara, intermediária e escura. Foi obtida uma sensibilidade de 98,4% para peles claras, 73,7% para intermediários e 92,1% para peles escuras.

Já existem alguns estudos publicados avaliando a performance do sistema *IrisPlex* em outras populações, sejam elas europeias (Itália, Portugal e Eslovênia) ou não (Ásia, Brasil, EUA e Venezuela) (Kastelic et al. 2013; Dembinski e Picard 2014; Freire-Aradas et al. 2014; Yun et al. 2014; Dario et al. 2015; Salvoro et al. 2019). A avaliação do sistema *IrisPlex* é especialmente importante em populações miscigenadas pois essa característica leva à uma grande incidência de fenótipos intermediários (por exemplo, olhos verdes são mais comuns que olhos azuis em nossa amostra). Com exceção do estudo realizado em Portugal, no qual um novo modelo de regressão logística criado pelo autor foi utilizado, resultados similares foram obtidos nos demais: a ferramenta não conseguiu prever olhos intermediários para nenhum indivíduo (sensibilidade = 0). Em um destes estudos, os modelos preditivos *Snipper* também foram analisados em amostras da população brasileira ($n = 79$) e venezuelana ($n = 20$) e se mostraram eficientes na predição de olhos azuis (sensibilidade variando de 0,879 a 0,946) e castanhos (sensibilidade variando de 0,844 a 0,941). Os sistemas conseguiram prever olhos

intermediários, porém com baixa sensibilidade (variando de 0,125 a 0,316) (Freire-Aradas et al. 2014).

Apesar de já existirem alguns estudos analisando a acurácia do sistema *IrisPlex* e *Snipper* para predição de cor de olhos em várias populações, nenhum estudo ainda foi realizado com o âmbito de analisar os sistemas *HIrisPlex* e *Snipper12* (cor de cabelo), *HIrisPlex-S* e *Snipper10* (cor de cabelo e pele) em populações que não aquelas utilizadas para o desenvolvimento dos sistemas.

1.3 POPULAÇÃO BRASILEIRA

A população brasileira pode ser considerada como uma das mais heterogêneas do mundo (Santos et al. 2010). A colonização do Brasil tem seu início no século XVI, quando aproximadamente 500 mil portugueses (homens, em sua maioria) chegaram em uma terra antes habitada por uma população de aproximadamente 2,5 milhões de indígenas. A miscigenação brasileira se deu início então com uma população composta principalmente por homens portugueses e mulheres nativas ameríndias (Pena et al. 2011). Em adição a isso, entre os séculos XVI e XIX estima-se que 3,5 milhões de africanos foram trazidos como escravos do oeste, centro-oeste e sudeste da África para o Brasil (Ferreira et al. 2006).

Posteriormente, o Brasil entrou em um período em que o governo desejava que houvesse um “embranquecimento” da população, estimulando a imigração de europeus para o país. Assim, estima-se que entre os anos de 1872–1975 houve uma intensa onda imigratória para o Brasil, principalmente nas regiões Sul e Sudeste, com a vinda de aproximadamente 5,4 milhões de indivíduos da Europa (em sua maioria italianos e portugueses) e do Oriente Médio (Santos et al. 2010; Pena et al. 2011). É importante lembrar que no século XIX a população ameríndia já havia diminuído consideravelmente, devido a guerras e doenças trazidas pelos europeus. Assim, a contribuição ameríndia se mostrou muito influente até meados de 1700, porém com a contínua vinda dos africanos e europeus para o Brasil, a mesma foi se reduzindo ao longo dos anos (Pena et al. 2011).

Dessa forma, a população brasileira atual é originada por uma mistura inter étnica envolvendo principalmente três grupos parentais: europeus, africanos e ameríndios, resultando

em uma grande variabilidade de cores de pele, indo de muito claro a muito escuro numa escala de tons contínuos (Santos et al. 2010; Pena et al. 2011).

Como a imigração dos povos se deu de forma irregular em nosso país, a miscigenação também se originou de diferentes maneiras, possuindo diferentes proporções de mistura nas cinco regiões geopolíticas do Brasil: no Nordeste, tem-se uma alta contribuição africana, enquanto o Norte apresenta uma contribuição ameríndia mais forte e o Sul e Sudeste, uma forte influência europeia (Salzano e Sans 2014).

Os resultados do censo realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) refletem os diferentes padrões da miscigenação brasileira. A nível nacional, 47,73% dos entrevistados consideraram-se brancos, 7,61% consideraram-se pretos, 1,09% consideraram-se amarelos, 43,13% consideraram-se pardos e 0,43%, indígenas. Porém, essas proporções variam de região para região: por exemplo, no Nordeste apenas 29,4% consideraram-se brancos, enquanto na região Sul do país 78,5% consideraram-se brancos. No estado de São Paulo, os dados do censo revelaram que 63,91% dos entrevistados se autodeclararam brancos, 5,52% consideraram-se pretos, 1,35% consideraram-se amarelos, 29,11% consideraram-se pardos, 0,10% consideraram-se indígenas e 0,01% não souberam se classificar. Por fim, na mesorregião de Ribeirão Preto, local onde nossas amostras foram coletadas, 65,94% dos entrevistados se autodeclararam brancos, 5,93% consideraram-se pretos, 0,72% consideraram-se amarelos, 27,33% consideraram-se pardos, 0,07% consideraram-se indígenas e 0,01% não souberam se classificar.

Como citado anteriormente, o Brasil possui uma variação muito grande de cores de pele, seguindo uma escala de tons contínuos. Assim, não existe um limiar definido para classificar uma pessoa de acordo com seu tom de pele, o que pode confundir um indivíduo em relação a sua autoclassificação: pessoas que se identificam como brancas ou pretas normalmente são “relativamente brancas” ou “relativamente pretas”, ou seja, uma pessoa que se classifica como parda pode ter um tom de pele muito próximo de outro indivíduo que se classifique como preto. O que define isso muitas vezes é o ambiente que a pessoa vive ou como ela se vê no contexto sócio-cultural/econômico do país. Além disso, o período da escravidão desencadeou um evidente racismo na população que se estende até hoje, fazendo com que as pessoas por muitas vezes queiram se classificar como mais claras do que realmente são (Telles 1992).

Vários estudos já mostraram que nem sempre a ancestralidade estará fortemente relacionada com a cor da pele (Parra et al. 2003; Pimenta et al. 2006; Pena et al. 2009). Ainda

assim, a percepção da ancestralidade de um indivíduo é normalmente estabelecida a partir da cor da pele, levando-se também em consideração outras características fenotípicas, como a cor dos olhos e cabelo e traços faciais. Por exemplo, um indivíduo com olhos azuis e cabelos lisos e loiros pode aparentar maior ancestralidade europeia do que realmente possui, enquanto um indivíduo de pele escura, cabelos crespos e lábios grossos pode aparentar ancestralidade africana muito maior do que a real. Além dos traços fenotípicos, fatores socioeconômicos e nível de educação também podem influenciar na auto avaliação de ancestralidade (Rodrigues de Moura et al. 2015).

Um estudo compilou os resultados de várias outras pesquisas avaliando a ancestralidade tri-híbrida brasileira publicadas anteriormente, resultando em uma população brasileira com contribuições europeia, africana e ameríndia de 0,62, 0,21 e 0,17, respectivamente. A região sul apresentou a maior contribuição europeia (0,77), enquanto o nordeste apresentou a maior contribuição africana (0,27) e o norte maior contribuição ameríndia (0,32) (Rodrigues de Moura et al. 2015). Um estudo recente (Valle-Silva et al. 2019) avaliou a contribuição média de cada grupo parental em uma amostra populacional de Ribeirão Preto utilizando dados de SNPs autossômicos, apresentando as seguintes contribuições: 0,72 a 0,79 europeia, 0,16 a 0,22 africana e 0,04 a 0,06 ameríndia. Esse mesmo estudo reanalisou os dados de outros dois estudos (Ferreira et al. 2006; Muniz et al. 2008) que avaliaram a ancestralidade de outro conjunto de indivíduos da região de Ribeirão Preto, estratificados em diferentes cores de pele, obtendo valores totais bastante condizentes: 0,72 europeia, 0,20 africana e 0,09 ameríndia.

Como já mencionado anteriormente, a alta miscigenação da população brasileira dá origem a uma variabilidade fenotípica muito grande, o que não é tão observado em populações parentais não-miscigenadas. Essa característica faz com que a população brasileira apresente condições propícias para o estudo da pigmentação humana, possibilitando a descoberta de novos alelos ou haplótipos que estejam de alguma forma influenciando na via de biossíntese da melanina e, por conseguinte, na determinação de características fenotípicas de pigmentação.

Justificativa

2. Justificativa

A aplicabilidade e acurácia das ferramentas para predição de cor de pele, bem como os modelos estatísticos implementados no *Snipper*, até agora não foram avaliadas na população brasileira. Polimorfismos desconhecidos ou diferentes daqueles normalmente considerados, bem como interações gene-gene diferentes das que ocorrem em populações parentais, podem estar influenciando o surgimento de tais fenótipos em populações miscigenadas, o que pode reduzir a acurácia das análises preditivas realizadas por meio das ferramentas disponíveis. Apesar disso, por apresentarem características genético-demográficas que diferem daquelas populações que serviram de base para a criação e validação das ferramentas abordadas, testar estas ferramentas em amostras populacionais brasileiras é de suma importância para verificar a abrangência e confiabilidade geral de seu uso para a prática forense.

Hipótese

3. Hipótese

Tem-se por hipótese que os sistemas preditivos desenvolvidos com base em populações europeias não terão êxito ao serem testados em populações altamente miscigenadas, como a brasileira, falhando principalmente na predição de fenótipos intermediários.

Objetivos

4. Objetivos

4.1 OBJETIVO GERAL

Avaliar a eficiência de diferentes ferramentas e modelos estatísticos desenvolvidos para a predição da pigmentação da pele, olhos e/ou cabelos em amostra constituída por 501 indivíduos miscigenados da região de Ribeirão Preto.

4.2 OBJETIVOS ESPECÍFICOS

- Genotipar um conjunto de 61 marcadores genéticos voltados para a predição de fenótipos de pigmentação humana a partir de dois conjuntos de experimentos previamente conduzidos com a amostra populacional em questão: (a) sequenciamento envolvendo as regiões promotora e codificadora de um painel de 16 genes de pigmentação e/ou (b) genotipagem em larga escala por meio do *Infinium Multi-Ethnic Global Array* (MEGA Array).
- Imputar os genótipos dos 11 marcadores ausentes em tais ensaios com base em SNPs presentes no MEGA Array e em padrões globais de desequilíbrio de ligação.
- Avaliar a associação individual de todos os marcadores com características morfológicas (pele, cabelo e olhos).
- Avaliar a acurácia de todas as ferramentas de predição consideradas no presente estudo.
- Comparar a acurácia dos resultados obtidos com o modelo matemático preditivo desenvolvido pelos criadores da ferramenta *HirisPlex* com predições geradas pelo modelo *Snipper*.
- Verificar se os diferentes conjuntos de marcadores preditivos podem ser úteis na inferência da ancestralidade individual.

Material e Métodos

5. Material e Métodos

5.1 COLETA

Amostras de sangue de 501 indivíduos de sexo feminino ($n = 252$) e masculino ($n = 249$), entre 18 e 70 anos, foram coletadas principalmente no Hemocentro de Ribeirão Preto – SP, e em outras dependências do *campus* da USP de Ribeirão Preto. Os 10 mL de sangue fornecidos pelos voluntários foram coletados com o auxílio da equipe de enfermagem do Hemocentro de Ribeirão Preto e técnicos especializados do Departamento de Genética (Laboratório de Genética Bioquímica) da FMRP/USP e armazenados em Vacutainer® com anticoagulante EDTA a 4°C até a etapa de extração.

Conforme proposto no projeto aprovado pelo Comitê de Ética em Pesquisa da FFCLRP/USP (Processo CEP – FFCLRP – CAAE n.25696413.7.0000.5407), os doadores assinaram um Termo de Consentimento Livre e Esclarecido (Anexos A e B), receberam uma cópia assinada pelos pesquisadores responsáveis e responderam a um questionário no qual se classificaram em relação à sua cor de pele (de acordo com a classificação de Fitzpatrick (1988) e a classificação do IBGE), cabelos e olhos, gerando assim informações de autoclassificação. Dois integrantes do Laboratório de Pesquisas Forenses e Genômicas – LPFG (nosso grupo de pesquisa) também classificaram os fenótipos de pele, olhos e cabelo do doador, gerando informações de heteroclassificação.

Medidas espectrofotométricas de áreas de baixa e alta exposição solar também foram coletadas em triplicata para mensurar o índice de melanina por meio do espectrofotômetro portátil CM-2500d e do *software SkinAnalysis* (Konica Minolta Holdings, Inc.).

Os questionários, fotos e medidas espectrofotométricas foram armazenados para posterior confronto com os dados obtidos no sequenciamento das amostras.

5.2 PROCEDIMENTOS LABORATORIAIS

5.2.1 Extração (Anexo C – Item 1)

O DNA das amostras coletadas foi extraído a partir do sangue total pela metodologia *salting-out* (Miller et al. 1988), com modificações de acordo com o protocolo descrito no Anexo C – item 1. O DNA foi ressuspenso em 200µL de água e armazenado a -20°C.

5.2.2 Quantificação (Anexo C – Item 2)

Após a extração, o DNA foi diluído em água em proporção 1:10 e quantificado por meio de ensaios espectrofotométricos (*NanoDrop™ 1000 Spectrophotometer - Thermo Scientific*), a fim de determinar a possível presença de proteínas contaminantes na amostra, e fluorimétricos (*Qubit - Life Technologies*), para a determinação da concentração do DNA genômico. Todas as amostras também foram submetidas a eletroforese em gel de agarose, com o intuito de verificar eventual degradação no DNA. Por fim, o DNA genômico foi diluído até atingir uma concentração final de 5 ng/µL.

5.2.3 Preparo e validação das bibliotecas (Anexo C – Itens 3 a 6)

As bibliotecas de fragmentos foram preparadas por meio de ensaio customizado implementado no kit *HaloPlex Target Enrichment System (Agilent Technologies, Inc.)*, com sondas desenhadas a partir de um ensaio *in silico (SureDesign, Agilent Technologies, Inc.)*, visando a captura das regiões regulatórias e exônicas de 16 genes envolvidos na biossíntese de melanina, bem como pequenas regiões contendo outros polimorfismos genéticos de interesse forense.

O preparo das bibliotecas foi feito de acordo com o protocolo *HaloPlex Target Enrichment System For Illumina Sequencing (Agilent Technologies, 2013)* com algumas alterações detalhadas nos itens 3 a 5 do Anexo C. Resumidamente, o processo consiste em cinco etapas:

a) Digestão das amostras por enzimas de restrição;

- b) Hibridização dos fragmentos alvo com sondas bionitiladas associadas a índices para identificação posterior de cada indivíduo;
- c) Captura dos fragmentos hibridizados utilizando *beads* magnéticas recobertas por estreptavidina;
- d) Amplificação dos fragmentos alvo por PCR;
- e) Purificação das bibliotecas.

As etapas de validação e quantificação das bibliotecas foram realizadas por eletroforese capilar *2100 Bioanalyzer (Agilent Technologies, Inc.)* e fluorimetria (*Qubit® 2.0 Fluorometer, Thermo Fisher Scientific Inc.*), especificados no item 6 do Anexo C. Após a quantificação as bibliotecas foram armazenadas a -20°C.

5.2.4 Sequenciamento de nova geração (NGS) (Anexo C – Item 7)

Após a validação, até 96 bibliotecas foram combinadas em um *pool* diluído de forma a obter 600 µL de uma solução com concentração final de 16 pM, na qual todas as bibliotecas estão, em princípio, presentes na mesma proporção, tendo assim representatividades parecidas. É muito importante que se observe a concentração final do *pool*, pois concentrações maiores do que a recomendada podem ocasionar uma alta clusterização, impedindo a interpretação do sinal, enquanto que concentrações menores podem não atingir a sensibilidade mínima do equipamento.

Essa solução foi então inserida no cartucho *MiSeq Reagent Kit V3 (600 cycles)*, de acordo com as recomendações do fabricante (item 7 do Anexo C), para sequenciamento *paired-end* na plataforma de sequenciamento *MiSeq Personal Sequencer (Illumina Inc.)*.

5.2.5 Genotipagem em larga escala por Array

As mesmas amostras de DNA foram submetidas a ensaio de genotipagem de aproximadamente 1,8 milhão de SNPs utilizando o chip *Infinium Multi-Ethnic Global-8 (MEGA) Kit (Illumina)* e a plataforma *HiScanSQ System (Illumina Inc.)*. Este ensaio de

genotipagem foi realizado no Centro de Genômica Funcional da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP), em parceria com os professores doutores Esteban J. Parra, da Universidade de Toronto, e Heather L. Norton, da Universidade de Cincinnati. O procedimento foi realizado seguindo as recomendações do fabricante, sem nenhuma alteração. O protocolo está disponível em https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium_lcg_assay/infinium-lcg-assay-reference-guide-15023139-04.pdf.

Resumidamente, o processo consiste em seis etapas:

- a) Amplificação do DNA genômico;
- b) Fragmentação enzimática do DNA;
- c) Precipitação e ressuspensão do DNA;
- d) Distribuição e hibridização do DNA com os *primers* contidos na *BeadChip*;
- e) Extensão do *primer* em uma única base marcada e identificação do sinal emitido;
- f) Leitura da *BeadChip*, a fim de coletar dados de intensidade da coloração.

5.3 PROCEDIMENTOS COMPUTACIONAIS (Anexo D)

No protocolo encontrado no Anexo D podem ser encontrados os procedimentos realizados para tratamento e análise de dados gerados por NGS descritos de maneira detalhada. Nos próximos tópicos cada passo será brevemente apresentado.

5.3.1 Trimagem e mapeamento dos dados obtidos por NGS

Durante o sequenciamento, nucleotídeos modificados, marcados com fluoróforos, vão sendo adicionados à sequência de cada *read*, gerando sinais que são captados pelo detector. Esses sinais são armazenados computacionalmente na forma de arquivos FastQ, podendo ser

obtidos acessando o próprio sequenciador ou a plataforma *Basespace* (<https://basespace.illumina.com>).

Estes arquivos FastQ gerados pelo sequenciamento armazenam as *reads* sequenciadas e as informações de qualidade (*Q-score*) para cada uma das posições de cada *read*. Para prosseguir com o tratamento de dados, os adaptadores adicionados durante o preparo das bibliotecas devem ser removidos (trimados), assim como as sequências de baixa qualidade. Para isso, a ferramenta *cutadapt* foi utilizada (Martin 2011).

Após a trimagem, as *reads* obtidas foram mapeadas e alinhadas ao genoma de referência hg19 utilizando o algoritmo BWA MEM (Li e Durbin 2009). O arquivo SAM (*.sam) obtido foi convertido em um arquivo BAM (*.bam) utilizando o programa SAMtools (<http://samtools.sourceforge.net/>).

5.3.2 Obtenção de genótipos

A partir dos dois conjuntos de dados genéticos foram recuperados dados genotípicos de um conjunto de 60 SNPs e 1 InDel pertencentes aos oito sistemas de predição de fenótipos de pigmentação considerados no presente estudo (Tabela 2): *IrisPlex*, *HirisPlex*, *HirisPlex-S*, *Snipper7*, *Snipper13*, *Snipper23*, *Snipper12*, e *Snipper10* (Walsh, Liu, et al. 2011; Ruiz et al. 2013; Walsh et al. 2013; Maroñas et al. 2014; Söchtig et al. 2015; Chaitanya et al. 2018).

Trinta e dois (32) marcadores foram obtidos por NGS utilizando a ferramenta *HaplotypeCaller* do pacote GATK (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera, Mauricio O. Carneiro, et al. 2013). Um arquivo vcf contendo os genótipos desses marcadores para 463 amostras (pois as outras 38 amostras possuíam muitos genótipos interrogados e tiveram seus dados de sequenciamento desconsiderados) foi gerado e este mesmo arquivo foi posteriormente processado pelo *software* VCFx (Castelli et al. 2015; Lima et al. 2016), com a finalidade de se avaliar a quantidade e qualidade das *reads* nos pontos de variação encontrados. Essa ferramenta retorna um arquivo vcf no qual os alelos pouco confiáveis, considerados como incertos, são substituídos por um ponto. A título de confirmação, os genótipos interrogados foram analisados visualmente utilizando o *Integrative Genomics Viewer* 2.3.75 (IGV) (Robinson et al. 2011; Thorvaldsdottir et al. 2013). Este *software* mostra as *reads* alinhadas a uma determinada região para determinado indivíduo. Assim, pode-se determinar se

o VCFx interrogou corretamente os pontos e realizar a correção, se necessário, de acordo com regras estabelecidas em função da quantidade e qualidade de *reads*, descritas no Anexo D.

Quarenta e dois (42) marcadores estavam presentes no ensaio de genotipagem realizado utilizando o chip MEGA Array. Os dados de genotipagem podem ser acessados por meio do *software GenomeStudio (Illumina)*, o qual também interroga os genótipos com baixa confiabilidade gerados pelo ensaio. As informações geradas foram exportadas por meio do *software plink v1.07* (Purcell et al. 2007), resultando em arquivos no formato ped e map, os quais posteriormente foram convertidos em um arquivo vcf. O manual do *GenomeStudio* está disponível em https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/genomestudio/genomestudio-2-0/genomestudio-genotyping-module-v2-user-guide-11319113-01.pdf.

Vinte e quatro (24) marcadores estavam presentes em ambos os ensaios de sequenciamento e genotipagem. Os dados obtidos por cada método foram comparados com o intuito de co-validar os experimentos e estimar taxas de discordância. Quando genótipos discordantes eram encontrados, considerava-se o genótipo em questão como interrogado.

Onze (11) marcadores não estavam presentes em nenhum dos ensaios laboratoriais realizados. Assim, ferramentas de imputação foram utilizadas a fim de determinar os genótipos com base em padrões globais de desequilíbrio de ligação, utilizando-se dados gerados pelo projeto *1000 Genomes* (The 1000 Genomes Consortium 2015). Nesta etapa, foram utilizados apenas os dados do MEGA Array. Duas ferramentas foram empregadas a fim de comparar os resultados: IMPUTE2 (Howie et al. 2009; Howie et al. 2011) e *Michigan Imputation Server* (MIS) (Das et al. 2016).

Como base de imputação para os programas IMPUTE2 e Michigan Imputation Server, os genótipos dos 100 marcadores presentes no Array mais próximos dos SNPs alvo (50 marcadores em posições anteriores e 50 em posições posteriores) foram utilizados. Apenas quatro regiões cromossômicas foram submetidas para imputação, pois 8 SNPs alvo do cromossomo 15 foram imputados em um único ensaio devido à sua proximidade.

Para imputação por meio do *software IMPUTE2* (Howie et al. 2009; Howie et al. 2011), primeiramente os genótipos obtidos pelo MEGA Array foram combinados com os dados de populações de referência (*1000 Genomes – Phase 3*) utilizando-se o *software SHAPEIT* (Delaneau et al. 2012; Delaneau et al. 2013). Nesta etapa, todos os sítios que não estão presentes nas populações de referência são retirados dos arquivos de entrada e, em cada ponto mantido,

os alelos alternativo e referência de cada marcador são verificados para confirmar se a mesma nomenclatura está sendo utilizada, ou então realizar os ajustes necessários. Como resultado um arquivo no formato bed é gerado. Utilizando esse arquivo e um intervalo de pares de base abrangendo a posição dos marcadores utilizados como base, o *software* IMPUTE2 realizou a imputação de todos os pontos contidos no intervalo, com exceção dos marcadores base utilizados, resultando em um arquivo no formato gen, o qual pode ser convertido em vcf utilizando o *software* gtool. O IMPUTE2 também gera um arquivo com informação sobre cada ponto (frequência dos alelos e qualidade da imputação).

Para imputação utilizando a ferramenta online *Michigan Imputation Server* (disponível em <https://imputationserver.sph.umich.edu/index.html>) (Das et al. 2016), os mesmos arquivos bed gerados pelo *software* SHAPEIT (Delaneau et al. 2012; Delaneau et al. 2013) foram convertidos em arquivos vcf utilizando o *software* plink (Purcell et al. 2007). Desses arquivos vcf foram extraídos os genótipos presentes no intervalo de imputação estabelecido utilizando o *software* vcftools (Danecek et al. 2011). Esse arquivo foi compactado em formato gz utilizando o *software* bgzip 1.9 e submetido ao servidor online do *Michigan Imputation Server*. O programa gera como resultado um arquivo vcf compactado, além de um arquivo de texto com as informações sobre a imputação em cada ponto (frequência dos alelos e qualidade da imputação).

Os intervalos de imputação escolhidos para cada marcador estão contidos na Tabela 1. Os genótipos obtidos pelos dois métodos foram aceitos como verdadeiros e utilizados apenas se concordassem entre si.

Tabela 1 – Regiões imputadas pelos *softwares* IMPUTE2 e *Michigan Imputation Server*.

Cromossomo	Intervalo de Imputação	Marcador(es) alvo(s)
12	88786155-89195185	rs10777129
15	28202805-29225935	rs2238289; rs6497292; rs7183877; rs8024968; rs12592730; rs12441727; rs4778241; rs1375164
16	89347441-89499232	rs3114908
16	89789691-89894033	rs12931267

Além disso, para verificar a confiabilidade da qualidade fornecida pelos *softwares*, genótipos de outros 11 marcadores presentes apenas no ensaio por NGS foram imputados

(imputação teste), a fim de calcular a porcentagem de acerto de cada ferramenta comparando os dados obtidos por meio da imputação e os dados reais de NGS.

Todos os 61 SNPs alvo contidos no projeto e a forma de obtenção de seu genótipo estão descritos na Tabela 2.

Tabela 2 – SNPs estudados, incluindo informações referentes à localização cromossômica e gênica, meio de obtenção dos genótipos de cada indivíduo e distribuição em painéis de SNPs para predição de fenótipos de pigmentação.

SNP	Localização (hg19)	Gene	Fonte de origem do dado genotípico	Painéis de SNPs
rs16891982	Chr5:33951693	SLC45A2	Ensaio NGS	A, B, C, D, E, F, H
rs28777	Chr5:33958959	SLC45A2	Ensaio NGS/MEGA	B, C, G
rs26722	Chr5:33963870	SLC45A2	Ensaio NGS/MEGA	F
rs13289	Chr5:33986409	SLC45A2	Ensaio NGS	H
rs12203592	Chr6:396321	IRF4	Ensaio NGS/MEGA	A, B, C, D, E, F
rs4959270	Chr6:457748	EXOC2	Ensaio NGS/MEGA	B
rs683	Chr9:12709305	TYRP1	Ensaio NGS/MEGA	B, C, F
rs1408799	Chr9:12672097	TYRP1	Ensaio MEGA	F, H
rs10756819	Chr9:16858084	BNC2	Ensaio MEGA	C
rs1393350	Chr11:89011046	TYR	Ensaio NGS/MEGA	A, C, D, E, F
rs1042602	Chr11:88911696	TYR	Ensaio NGS/MEGA	B, C
rs1126809	Chr11:89017961	TYR	Ensaio NGS	C
rs35264875	Chr11:68846399	TPCN2	Ensaio MEGA	G
rs3829241	Chr11:68855363	TPCN2	Ensaio MEGA	H
rs12821256	Chr12:89328335	KITLG	Ensaio NGS/MEGA	B, C
rs10777129	Chr12:88961713	KITLG	Imputação	H
rs12896399	Chr14:92773663	SLC24A4	Ensaio NGS/MEGA	A, C, D, E, F
rs2402130	Chr14:92801203	SLC24A4	Ensaio NGS/MEGA	B, C, H
rs17128291	Chr14:92882826	SLC24A4	Ensaio MEGA	C
rs1426654	Chr15:48426484	SLC24A5	Ensaio MEGA	C, H
rs12913832	Chr15:28365618	HERC2	Ensaio NGS/MEGA	A, B, C, D, E, F, G
rs2238289	Chr15:28453215	HERC2	Imputação	C
rs6497292	Chr15:28496195	HERC2	Imputação	C
rs1129038	Chr15:28356859	HERC2	Ensaio NGS/MEGA	C, D, E, F, G
rs1667394	Chr15:28530182	HERC2	Ensaio MEGA	C, E, F
rs11636232	Chr15:28386626	HERC2	Ensaio NGS	E, F
rs7183877	Chr15:28365733	HERC2	Imputação	E, F
rs8024968	Chr15:28283689	HERC2	Imputação	E, F
rs12592730	Chr15:28530359	HERC2	Imputação	F
rs916977	Chr15:28513364	HERC2	Ensaio MEGA	F
rs1800407	Chr15:28230318	OCA2	Ensaio NGS/MEGA	A, B, C, D, E, F
rs1800414	Chr15:28197037	OCA2	Ensaio NGS/MEGA	C
rs1470608	Chr15:28288121	OCA2	Ensaio MEGA	C
rs1545397	Chr15:28187772	OCA2	Ensaio MEGA	C
rs12441727	Chr15:28271775	OCA2	Imputação	C
rs4778232	Chr15:28281765	OCA2	Ensaio MEGA	E, F
rs4778241	Chr15:28338713	OCA2	Imputação	E, F

SNP	Localização (hg19)	Gene	Fonte de origem do dado genotípico	Painéis de SNPs
rs1375164	Chr15:28291812	OCA2	Imputação	F
rs4778138	Chr15:28335820	OCA2	Ensaio MEGA	F, G
rs7495174	Chr15:28344238	OCA2	Ensaio NGS/MEGA	F, G
rs1448484	Chr15:28283441	OCA2	Ensaio MEGA	H
rs885479	Chr16:89986154	MC1R	Ensaio NGS	B, C
rs796296176	Chr16:89985750	MC1R	Ensaio NGS	B
rs1805005	Chr16:89985844	MC1R	Ensaio NGS/MEGA	B
rs1805006	Chr16:89985918	MC1R	Ensaio NGS/MEGA	B, C, G
rs1805007	Chr16:89986117	MC1R	Ensaio NGS/MEGA	B, C, G
rs1805008	Chr16:89986144	MC1R	Ensaio NGS/MEGA	B, C, G
rs1805009	Chr16:89986546	MC1R	Ensaio NGS/MEGA	B, G
rs2228479	Chr16:89985940	MC1R	Ensaio NGS/MEGA	B, C
rs11547464	Chr16:89986091	MC1R	Ensaio NGS/MEGA	B, C, G
rs1110400	Chr16:89986130	MC1R	Ensaio NGS/MEGA	B, C
rs201353883	Chr16:89986122	MC1R	Ensaio MEGA	B
rs3212355	Chr16:89984378	MC1R	Ensaio NGS	C
rs3114908	Chr16:89383725	ANKRD11	Imputação	C
rs8051733	Chr16:90024206	DEF8	Ensaio MEGA	C
rs12931267	Chr16:89818732	FANCA	Imputação	G
rs2378249	Chr20:33218090	ASIP	Ensaio NGS/MEGA	B, C
rs6119471	Chr20:32785212	ASIP	Ensaio MEGA	C, H
rs1015362	Chr20:32738612	ASIP	Ensaio MEGA	F
rs6058017	Chr20:32856998	ASIP	Ensaio NGS	F, H
rs6059655	Chr20:32665748	RALY	Ensaio MEGA	C

Obs: A = *IrisPlex*; B = *HIrisPlex*; C = *HIrisPlex-S*; D = *Snipper7*; E = *Snipper13*; F = *Snipper23*; G = *Snipper12*; H = *Snipper10*.

5.3.3 Análises populacionais

As frequências alélicas, heterozigose observada e esperada e probabilidade de aderência ao equilíbrio de Hardy-Weinberg foram calculadas utilizando o *software* GenePop 4.51 (Rousset 2008). Para essa última análise, foram utilizados apenas os 338 indivíduos amostrados aleatoriamente.

Para identificar associações entre alelos e fenótipos de pigmentação foi realizada uma regressão logística utilizando o *software* plink v1.07 (Purcell et al. 2007). As associações encontradas foram avaliadas por meio do Teste Exato de Fisher bi-caudal e valores de *Odds Ratio* (OR), utilizando um Intervalo de Confiança (IC) de 95%. Quando possuindo valor de $p < 0,05$, $OR \geq 5$ e limite inferior do IC 95% > 1 , as associações foram consideradas como significativas, sendo classificadas como moderadas quando $5 \leq OR < 10$ e fortes quando $OR \geq 10$. Com essa análise é possível adicionar a ancestralidade de cada indivíduo como fator

de confusão, confirmando assim que as associações encontradas se referem a fenótipos de pigmentação, não possuindo nenhum viés da ancestralidade individual. Os dados de ancestralidade individual já foram previamente preparados por outros membros de nosso grupo de pesquisa (Oliveira et al. 2018) utilizando o conjunto de 34 marcadores SNPforID 34-plex (Phillips et al. 2007; Fondevila et al. 2013). Ao adicionar a ancestralidade como fator de confusão, todas as associações com $OR > 1$ foram consideradas.

5.3.4 Análises de predições fenotípicas

As amostras foram estratificadas em grupos de acordo com a pigmentação dos olhos, pele e cabelos. A cor dos olhos foi caracterizada e dividida em categorias compreendendo azul ($n = 44$), verde ($n = 81$), mel ($n = 22$), castanho-claro ($n = 125$), castanho-escuro ($n = 229$). A cor dos cabelos foi caracterizada e dividida em categorias compreendendo ruivo ($n = 12$), loiro-claro ($n = 18$), loiro-escuro ($n = 45$), castanho-claro ($n = 87$), castanho-escuro ($n = 175$) e preto ($n = 125$). Trinta e nove (39) indivíduos não puderam ter sua cor de cabelo identificada, pois tingiam o cabelo, possuíam cabelos grisalhos ou eram carecas. Já para a pele, os indivíduos foram classificados de duas maneiras: (a) de acordo com a classificação de Fitzpatrick (1988), a qual se baseia nas características da pele associadas à reação da mesma à exposição ao sol, resultando na determinação do tipo de pele em seis grupos: I ($n = 67$), II ($n = 136$), III ($n = 141$), IV ($n = 81$), V ($n = 49$), VI ($n = 25$), com dois indivíduos não classificados pois possuíam vitiligo em estado avançado, impossibilitando a determinação de sua cor de pele, e (b) de acordo com o valor L^*_{ab} obtido por meio do espectrofotômetro portátil CM2500D (Konica Minolta Holdings, Inc.), estratificando a pele dos indivíduos em branca ($n = 347$), intermediária ($n = 40$) e escura ($n = 6$), com oito indivíduos não classificados pois não foi possível obter valores de L^*_{ab} .

Para as categorias fenotípicas (cada cor considerada) pertencentes a cada ferramenta (conjuntos *Snipper* e *HIrisPlex-S*) foram calculados os parâmetros AUC (*Area Under the Receiver Operating Characteristic Curve*), sensibilidade, especificidade, valor preditivo positivo (PPV) e valor preditivo negativo (NPV). No caso do *HIrisPlex-S* a acurácia geral da ferramenta também foi considerada, isto é, a taxa de acerto combinada das predições para olho, cabelo e pele. A construção da curva ROC (*Receiver Operating Characteristic Curve*) e o cálculo da AUC associada foram realizados utilizando o *software* SPSS 20.0 (IBM). Enquanto

a curva ROC consiste em um gráfico que plota sensibilidade *versus* o complemento da especificidade (1-especificidade), a AUC associada consiste em um parâmetro que avalia a acurácia de ferramentas classificatórias, e nada mais é do que a integral da curva ROC, em que valores de 0 a 0,5 refletem total ausência de poder preditivo e o valor 1,0 representa uma predição perfeita realizada pela ferramenta (Söchtig et al. 2015). A Figura 3 ilustra três exemplos de curva ROC: a curva verde ilustra uma situação ideal, em que a ferramenta conseguiria prever perfeitamente todos os indivíduos, atingindo valores de sensibilidade e especificidade iguais a 1. Por consequência, a AUC associada também corresponde a 1. Já a curva azul ilustra uma situação satisfatória, na qual a ferramenta consegue prever corretamente alguns indivíduos atingindo bons níveis de sensibilidade e especificidade. Sua AUC associada assumiria valores próximos ou maiores que 0,8. Por fim, a curva vermelha ilustra a situação de uma ferramenta sem poder de predição, ou seja, que realiza predições aleatórias, não atingindo um equilíbrio satisfatório entre sensibilidade e especificidade, com AUC em torno de 0,5.

Figura 3 – Três curvas ROC de diferentes classificadores preditivos. A curva verde corresponde à uma ferramenta com predição perfeita, enquanto a curva azul corresponde à uma ferramenta com predição boa e a curva vermelha à uma ferramenta com poder de predição aleatório. Retirado de <http://crsouza.com/2009/07/13/analise-de-poder-discriminativo-atraves-de-curvas-roc/>.

Para melhor entender os outros parâmetros, usaremos o quadro ilustrado na Figura 4. Como exemplo, usaremos a cor de olho azul. Os resultados positivos são aqueles em que a ferramenta indica olho azul para um determinado indivíduo, enquanto os resultados negativos são aqueles nos quais a ferramenta indica uma cor de olho não azul (intermediário ou castanho) para determinados indivíduos. Os indivíduos com fenótipo presente são aqueles que realmente possuem olhos azuis e os indivíduos com fenótipos ausentes são aqueles que não possuem olhos azuis. A sensibilidade (dada por $a/(a+c)$) pode ser considerada como a taxa de sucesso de classificação da ferramenta. Ela expressa a proporção de indivíduos que apresentam um determinado fenótipo classificados corretamente; no presente exemplo, a proporção de indivíduos com olhos azuis que tiveram resultado positivo. Já a especificidade (dada por $d/(b+d)$) expressa a proporção de indivíduos que não apresentam o fenótipo em questão classificados corretamente; no presente exemplo, a proporção de indivíduos com olhos intermediários ou castanhos que tiveram resultado negativo para olho azul. O PPV (dado por $a/(a+b)$) expressa a proporção dos indivíduos classificados com um determinado fenótipo pela ferramenta que realmente apresentam essa característica; no presente exemplo, a proporção de indivíduos com resultado positivo que realmente possuem olhos azuis. Por fim, o NPV (dado por $d/(c+d)$) expressa a proporção de indivíduos classificados como não possuindo um determinado fenótipo pela ferramenta que realmente não o apresentavam; no presente exemplo, a proporção de indivíduos com resultado negativo para olho azul que realmente possuem olhos intermediários ou castanhos (Ruiz et al. 2013).

	Fenótipo Presente	Fenótipo Ausente	Total
Resultado Positivo	a	b	a+b
Resultado Negativo	c	d	c+d
Total	a+c	b+d	a+b+c+d

Figura 4 – Quadro delineando a distribuição dos resultados gerados por uma ferramenta preditiva na análise de um fenótipo binário.

5.3.4.1 *HirisPlex-S*

A predição de fenótipos a partir dos marcadores presentes no sistema *HirisPlex-S* é feita por meio de uma regressão logística multinominal (MLR) e pode ser realizada utilizando-se a ferramenta online e gratuita disponibilizada em <https://hirisplex.erasmusmc.nl/>. Nessa

plataforma, o usuário envia um arquivo csv contendo os genótipos de cada uma das amostras para o conjunto de 41 marcadores que compõem a ferramenta. A partir desses dados, a ferramenta calcula as probabilidades associadas a cada categoria de cada característica fenotípica: olhos (azul, intermediário e castanho), cabelos (cor: loiro, castanho, ruivo e preto; tom: claro ou escuro) e pele (muito clara, clara, intermediária, escura e muito escura). No caso dos olhos, os parâmetros foram calculados em três níveis de probabilidade (*p-value*): considerando que a cor com maior *p* corresponde ao fenótipo do indivíduo em questão e também aplicando *threshold* de 0,5 e 0,7. Nestes dois últimos cenários, indivíduos com valores de *p* abaixo do *threshold* foram considerados como “não-classificados”. No caso da cor do cabelo e pele, para se determinar a predição final deve-se seguir um guia com instruções de como interpretar os *p-values* gerados pela ferramenta (Walsh et al. 2014; Chaitanya et al. 2018).

Ainda em relação à cor dos cabelos, apesar de serem gerados *p-values* correspondentes às quatro cores mencionadas acima (loiro, castanho, ruivo e preto), ao seguir o guia de instruções, a predição pode resultar em cabelos loiro-escuros/castanho-claros (considerados como uma só cor pela ferramenta devido à alta semelhança entre eles) e castanho-escuros. Assim, para fins de comparação com os valores disponibilizados no site da ferramenta, a AUC foi calculada para as quatro cores que possuem *p-values* gerados, enquanto que os outros parâmetros (sensibilidade, especificidade, PPV e NPV) foram calculados com base na predição de tons resultantes de acordo com o guia.

Para a predição da cor de pele com o sistema *HIrisPlex-S*, a estratificação dos indivíduos deve ser feita de acordo com a escala de Fitzpatrick (1988), com categorias de I a VI, sendo I a pele mais clara e VI a pele mais escura. Apesar de haver seis categorias nessa escala, existem apenas cinco na ferramenta: *Very Pale* (muito clara – pele tipo I), *Pale* (clara – pele tipo II), *Intermediate* (intermediária – peles tipo III e IV), *Dark* (escura – pele tipo V) e *Dark-Black* (muito escura – pele tipo VI). Os parâmetros sensibilidade, especificidade, PPV e NPV foram calculados com base na predição resultante de acordo com o guia.

Todas as análises foram realizadas utilizando a ferramenta *HIrisPlex-S*, porém para que não haja confusões, quando citado o termo “conjunto *IrisPlex*” nesse trabalho, o mesmo refere-se à predição da cor dos olhos. Quando citado o termo “conjunto *HIrisPlex*” nesse trabalho, o mesmo refere-se à predição da cor dos cabelos. E por fim, quando citado o termo “conjunto *HIrisPlex-S*” nesse trabalho, o mesmo refere-se à predição da cor da pele, sendo apenas uma questão de nomenclatura.

5.3.4.2 *Snipper*

A ferramenta online disponível em <http://mathgene.usc.es/Snipper/> foi utilizada para realizar a predição dos fenótipos por meio dos outros cinco conjuntos de marcadores vinculados ao *Snipper app suite version 2.5*. Para tanto, é necessário inserir os genótipos dos conjuntos de marcadores mencionados e selecionar adequadamente as opções de análise. Como resultado, o sistema disponibilizará probabilidades de ocorrência de cada uma das categorias. A plataforma permite três opções de classificação: o usuário pode escolher entre utilizar um classificador Bayesiano (em que a ferramenta assume que os marcadores estão em equilíbrio de Hardy-Weinberg), realizar uma regressão logística multinomial (MLR) ou utilizar uma metodologia baseada em distâncias genéticas. Neste trabalho, apenas as duas primeiras opções foram utilizadas, pois a primeira é sugerida pelos desenvolvedores da ferramenta e a segunda é o modelo utilizado pelo sistema *HIrisPlex-S*. Além disso, para os conjuntos de predição de cor dos olhos e cabelo (*Snipper7*, *Snipper13*, *Snipper23* e *Snipper12*), pode-se escolher entre considerar os marcadores rs1129038-12913832 na forma de genótipos ou em fase haplotípica. Neste trabalho, buscando definir qual o melhor método de classificação, as análises foram feitas para todas as opções disponíveis. Ao considerarmos o haplótipo, os mesmos foram inferidos computacionalmente utilizando o *software* PHASE 2.1.1 (Stephens et al. 2001; Stephens e Donnelly 2003).

O conjunto *Snipper7* é composto pelos 6 SNPs do sistema *IrisPlex* com adição do marcador rs1129038. Assim, como forma de análise alternativa, apenas os genótipos referentes aos 6 SNPs do *IrisPlex* foram inseridos na plataforma *Snipper*, a fim de comparar os modelos preditivos incorporados nas duas plataformas. O conjunto de marcadores *IrisPlex* quando analisado na plataforma *Snipper* foi aqui denominado *Snipper6*.

Por fim, para a predição da cor dos cabelos, pode-se escolher entre classificar os fenótipos em duas (claro e escuro) ou quatro (loiro, castanho, preto e ruivo) categorias. A segunda opção foi a escolhida no presente estudo.

Seguindo-se as recomendações prévias dos desenvolvedores de cada ferramenta (Ruiz et al. 2013; Maroñas et al. 2014; Söchtig et al. 2015), os resultados obtidos utilizando os conjuntos *Snipper* com o classificador Bayesiano foram analisados em dois níveis: a) considerando que a cor com a probabilidade mais alta corresponde ao fenótipo do indivíduo em questão e b) estabelecendo um *cut-off* de 3:1, ou seja, a probabilidade mais alta fornecida pela

ferramenta deve ser três vezes maior do que a segunda probabilidade mais alta para que o resultado seja considerado válido. Indivíduos que não atingiram esse *cut-off* foram considerados como “não-classificados”.

Os resultados obtidos em todas as análises de predição foram comparados com os fenótipos reais dos indivíduos participantes do projeto. Posteriormente, os parâmetros foram comparados entre as ferramentas que predizem a mesma característica morfológica para verificar quais apresentam melhor desempenho e se existe viabilidade de uso no Brasil.

5.3.5 Análises de Ancestralidade

As proporções de ancestralidades individual e populacional foram inferidas para os seis conjuntos de marcadores preditivos (*HIrisPlex-S*, *Snipper7*, *Snipper13*, *Snipper23*, *Snipper10* e *Snipper12*) utilizando o *software* STRUCTURE 2.3.4 (Pritchard et al. 2000; Hubisz et al. 2009). O modelo de mistura com frequências alélicas correlacionadas foi empregado, utilizando os seguintes parâmetros: 100.000 interações de *burn-in*, seguido de 100.000 interações de *Markov Chain Monte Carlo*; pré definição de três *clusters* ($K = 3$), cada um representando uma das três ancestralidades principais da população brasileira: europeia, africana e ameríndia; 100 corridas independentes.

Para as análises, foram utilizados dados de 14 populações da terceira fase do projeto *1000 Genomes* (Apêndice A). Como esse projeto não avaliou um grupo de ameríndios autóctones apropriados para representar a ancestralidade nativo-americana, foram utilizadas populações do leste asiático como substituto da população parental ameríndia nas análises.

Resultados

6. Resultados

6.1 GENOTIPAGEM

Pela técnica de NGS e com o uso da ferramenta *HaplotypeCaller*, foi obtido um arquivo vcf contendo 32 marcadores em 501 amostras. Dessas amostras, 28 tiveram mais da metade dos marcadores não genotipados. Além disso, posteriormente foi constatado que 10 amostras apresentavam evidências de mistura com DNA de outros indivíduos em sua composição. Deste modo, 38 amostras não puderam ser genotipadas por este método. Assim, 463 dos 501 indivíduos apresentaram dados por NGS. Os genótipos obtidos foram posteriormente analisados pelo *software* VCFx em função da cobertura e qualidade de sequenciamento. Sessenta e cinco genótipos, em um universo de 14.816 genótipos analisados, foram interrogados, ou seja, houve a substituição do alelo considerado incerto por um ponto. Após esses sítios serem verificados individualmente utilizando o *software* IGV 2.3.75, em 14 deles o segundo alelo foi determinado, enquanto outros 51 pontos permaneceram interrogados. Por fim, das 463 amostras, 436 tiveram perfis completos gerados por meio do NGS.

Por meio do ensaio de genotipagem MEGA Array, foram obtidos os genótipos de 42 marcadores para 501 amostras, todas apresentando boa qualidade de genotipagem. Dos 21.042 pontos obtidos, o *software* *GenomeStudio* não conseguiu definir o genótipo de 46 deles. Assim, das 501 amostras, 480 tiveram perfis completos gerados por meio do Array.

Como 24 marcadores estavam presentes em ambos os ensaios de sequenciamento e genotipagem, os dados obtidos independentemente pelos dois métodos foram comparados a fim de verificar a confiabilidade dos genótipos. Para essa comparação, foram utilizadas as 463 amostras que apresentaram os dados de NGS, resultando em 11.112 pontos para comparação. Destes, foram excluídos aqueles que continham genótipos interrogados, restando 11.031 pontos. Foram encontradas apenas 40 inconsistências (0,36%).

Para que as 38 amostras relatadas no primeiro parágrafo não fossem excluídas do estudo, os genótipos de sete SNPs que deveriam ter sido obtidos exclusivamente por NGS foram imputados pelas ferramentas IMPUTE2 e *Michigan Imputation Server* utilizando os dados de Array. Não foi possível imputar os genótipos do oitavo marcador, uma vez que se trata da única

InDel presente em todo o conjunto de marcadores analisados. Ainda assim, este procedimento permitiu com que todos os 501 indivíduos fossem mantidos para análises subsequentes.

6.2 IMPUTAÇÃO

Como previsto, 11 marcadores presentes nos sistemas preditivos foram imputados com sucesso. Antes de proceder com a imputação de tais marcadores, foi realizado um procedimento de validação da imputação considerando um outro conjunto de 11 marcadores, cujos genótipos já haviam sido definidos em 260 amostras por NGS. Os resultados da comparação entre os dados reais obtidos por NGS e os dados de imputação gerados pelos dois *softwares* diferentes estão apresentados na Tabela 3. Adicionalmente, a fim de restringir os dados apenas às imputações com elevada acurácia, a taxa de acurácia também foi determinada em um grupo definido como “Unificado”, composto exclusivamente pelos genótipos com resultados consensuais produzidos por ambos os métodos. Cada um dos 11 marcadores apresentou genótipos concordantes entre 241 e 260 amostras (média de 253 amostras genotipadas por marcador). Considerando todos os 2.786 genótipos concordantes, apenas 27 deles (0,97%) não estavam corretos, variando de zero (em seis marcadores) a 21 (rs1800411) inconsistências entre o genótipo unificado e o estabelecido por NGS.

Tabela 3 – Marcadores escolhidos para validação dos resultados, qualidade teórica da imputação fornecida pelos softwares e acurácia real da imputação mensurada pela taxa de concordância com os genótipos determinados por NGS. Os números entre parênteses representam a quantidade de indivíduos com genótipos indeterminados.

Marcador	Qualidade Teórica		Acurácia Real		
	MIS	IMPUTE2	MIS	IMPUTE2	MIS
rs61738394	0,943	0,998	0,996	1,000 (0,38%)	1,000 (1,15%)
rs28595435	0,887	0,986	0,988	0,968 (4,63%)	0,992 (6,95%)
rs74005643	0,959	0,977	1,000	0,992 (0,39%)	1,000 (1,17%)
rs8039195	0,946	0,988	0,935	0,996 (2,69%)	0,996 (7,31%)
rs369504551	0,863	1,000	1,000	1,000 (0,00%)	1,000 (0,00%)
rs869408	0,959	0,992	0,985	0,992 (2,31%)	0,992 (2,69%)
rs1800411	0,784	0,994	0,900	0,911 (1,15%)	0,916 (3,46%)
rs4778214	0,723	0,987	0,992	0,996 (3,46%)	1,000 (4,23%)
rs567843695	0,573	0,998	0,988	0,996 (0,83%)	0,996 (1,24%)
rs569842723	0,609	0,998	1,000	1,000 (0,00%)	1,000 (0,00%)
rs150719576	0,518	0,999	1,000	0,996 (0,00%)	1,000 (0,38%)
Média	0,943	0,992	0,980	0,986 (0,97%)	0,990 (2,60%)

Diante do sucesso do procedimento de validação metodológico, a qualidade geral da imputação de cada um dos 11 marcadores presentes nos sistemas preditivos, fornecida pelos *softwares* IMPUTE2 e *Michigan Imputation Server* (MIS), está apresentada na Tabela 4.

Tabela 4 – Marcadores imputados e a qualidade da imputação fornecida por cada ferramenta.

Marcador	Qualidade teórica da imputação	
	MIS	IMPUTE2
rs10777129	0,962	0,973
rs12441727	0,918	0,984
rs8024968	0,994	0,996
rs1375164	0,995	0,999
rs4778241	0,847	0,809
rs7183877	0,699	0,890
rs22382891	0,440	0,981
rs6497292	0,965	0,993
rs12592730	0,735	0,978
rs3114908	0,868	0,920
rs12931267	0,655	0,956
Média	0,825	0,953

Os resultados obtidos com uso das duas ferramentas foram comparados a fim de verificar se os mesmos estavam concordantes. Os dados de discordância entre os dois métodos, bem como a porcentagem de genótipos não definidos pelo IMPUTE2 e o número de amostras com genótipo concordantes (unificados), estão apresentados na Tabela 5. Para as análises seguintes, serão utilizados apenas os genótipos que se mostraram concordantes (genótipos unificados).

A Tabela 6 apresenta os dados referentes aos sete SNPs presentes no ensaio de NGS que precisaram ser imputados para 38 amostras utilizando os dados de Array. Assim como no caso dos 11 marcadores imputados em todo o conjunto amostral, serão utilizados nas análises seguintes apenas os genótipos que se mostraram concordantes (genótipos unificados).

Tabela 5 – Discordâncias encontradas nos resultados obtidos pelos dois métodos de imputação.

Marcador	Porcentagem de genótipos não definidos pelo IMPUTE2	Taxa de discordâncias entre MIS e IMPUTE2	Número de amostras com genótipos concordantes
rs10777129	4,19%	0,83%	476
rs12441727	1,40%	0,81%	490
rs8024968	0,40%	0,41%	497
rs1375164	0,20%	0,20%	499
rs4778241	35,73%	8,70%	295
rs7183877	9,38%	1,98%	446
rs22382891	2,99%	4,73%	463
rs6497292	0,80%	0,60%	494
rs12592730	1,80%	1,42%	485
rs3114908	13,37%	6,22%	407
rs12931267	0,20%	0,60%	497
Média	6,45%	2,41%	459

Tabela 6 – Número de genótipos não definidos pelo IMPUTE2, de discordâncias entre as duas ferramentas de imputação e de amostras com genótipos concordantes

Marcador	Número de genótipos não definidos pelo IMPUTE2	Número de discordâncias entre MIS e IMPUTE2	Número de amostras com genótipos concordantes
rs16891982	8	1	29
rs13289	2	0	36
rs1126809	0	0	38
rs11636232	5	3	30
rs3212355	1	0	37
rs885479	2	7	29
rs6058017	2	0	36

6.3 ANÁLISES POPULACIONAIS

Conforme mencionado anteriormente, visando obter uma gama variada de fenótipos em nossa amostragem, a mesma foi enriquecida, convidando indivíduos de fenótipos mais raros (ex: cabelos loiro-claro, olhos azuis, cabelos ruivos) a participar do estudo. Assim, das 501 amostras presentes no estudo, foram utilizadas nas análises populacionais apenas aquelas amostradas aleatoriamente, resultando em um subconjunto de 338 amostras.

A Tabela 7 mostra as frequências alélicas, heterozigoses observada e esperada e aderência ao Equilíbrio de Hardy-Weinberg (p_{EHW}) para os 61 pontos de variação.

Tabela 7 – Frequências alélicas, heterozigoses observada e esperada (H_O e H_E) e aderência ao Equilíbrio de Hardy-Weinberg (p_{EHW}) referentes aos 61 marcadores de estudo. Desvios do equilíbrio de Hardy-Weinberg estão destacados em negrito. Estão sublinhados os marcadores que não desviaram do EHW após a correção de Bonferroni ($p = 0,00082$). Os marcadores estão ordenados pela posição cromossômica que ocupam.

Marcador	n	H_O	H_E	Alelo referência (Frequência)	Alelo alternativo (Frequência)	p_{EHW}
rs16891982 ^a	334	0,419	0,489	C (0,422)	G (0,578)	<u>0,010</u>
rs28777 ^b	333	0,402	0,455	A (0,652)	C (0,348)	<u>0,040</u>
rs26722 ^b	337	0,214	0,219	C (0,875)	T (0,125)	0,622
rs13289 ^a	337	0,496	0,499	C (0,527)	G (0,473)	0,913
rs12203592 ^b	336	0,146	0,151	C (0,918)	T (0,082)	0,475
rs4959270 ^b	335	0,487	0,490	C (0,575)	A (0,425)	0,912
rs683 ^b	332	0,485	0,501	C (0,492)	A (0,508)	0,584
rs1408799 ^c	338	0,509	0,497	C (0,544)	T (0,456)	0,662
rs10756819 ^c	338	0,476	0,497	A (0,546)	G (0,454)	0,510
rs1393350 ^b	337	0,255	0,270	G (0,840)	A (0,160)	0,316
rs1042602 ^b	336	0,396	0,434	C (0,683)	A (0,317)	0,130
rs1126809 ^a	324	0,268	0,284	G (0,829)	A (0,171)	0,328
rs35264875 ^c	338	0,216	0,211	T (0,880)	A (0,120)	0,801
rs3829241 ^c	336	0,438	0,389	G (0,737)	A (0,263)	<u>0,024</u>
rs12821256 ^b	338	0,086	0,082	T (0,957)	C (0,043)	1,000
rs10777129 ^d	321	0,312	0,316	G (0,804)	A (0,196)	0,859
rs12896399 ^b	334	0,416	0,412	G (0,711)	T (0,289)	0,894
rs2402130 ^b	333	0,390	0,403	A (0,721)	G (0,279)	0,587
rs17128291 ^c	333	0,213	0,205	A (0,884)	G (0,116)	0,595
rs1426654 ^c	338	0,314	0,377	A (0,749)	G (0,251)	<u>0,004</u>
rs12913832 ^b	337	0,406	0,422	A (0,699)	G (0,301)	0,519
rs2238289 ^d	308	0,442	0,469	A (0,627)	G (0,373)	0,331
rs6497292 ^d	334	0,347	0,365	A (0,760)	G (0,240)	0,371
rs1129038 ^b	337	0,418	0,424	C (0,696)	T (0,304)	0,798
rs1667394 ^c	336	0,518	0,498	A (0,542)	G (0,458)	0,510
rs11636232 ^a	337	0,240	0,234	C (0,865)	T (0,135)	0,815
rs7183877 ^d	297	0,195	0,208	C (0,882)	A (0,118)	0,268
rs8024968 ^d	334	0,422	0,406	C (0,717)	T (0,283)	0,503
rs12592730 ^d	324	0,213	0,210	G (0,881)	A (0,119)	1,000

Marcador	<i>n</i>	H _O	H _E	Alelo referência (Frequência)	Alelo alternativo (Frequência)	<i>p</i> EHW
rs916977 ^c	338	0,497	0,494	G (0,556)	A (0,444)	1,000
rs1800407 ^b	338	0,118	0,122	C (0,935)	T (0,065)	0,641
rs1800414 ^b	338	0,009	0,015	T (0,993)	C (0,007)	<u>0,015</u>
rs1470608 ^c	337	0,487	0,479	C (0,605)	A (0,395)	0,820
rs1545397 ^c	338	0,210	0,216	A (0,877)	T (0,123)	0,614
rs12441727 ^d	328	0,296	0,282	C (0,831)	A (0,169)	0,434
rs4778232 ^c	337	0,493	0,497	G (0,543)	A (0,457)	0,913
rs4778241 ^d	187	0,524	0,388	A (0,738)	C (0,262)	0,000
rs1375164 ^d	337	0,493	0,499	C (0,528)	T (0,472)	0,828
rs4778138 ^c	338	0,488	0,469	A (0,626)	G (0,374)	0,487
rs7495174 ^b	338	0,337	0,307	A (0,811)	G (0,189)	0,078
rs1448484 ^c	332	0,235	0,256	A (0,849)	G (0,151)	0,135
rs885479 ^a	334	0,108	0,113	G (0,940)	A (0,060)	0,329
rs796296176 ^a	308	0,000	0,000	- (1,0000)	A (0,0000)	-
rs1805005 ^b	338	0,169	0,179	G (0,901)	T (0,099)	0,351
rs1805006 ^b	337	0,018	0,018	C (0,991)	A (0,009)	1,000
rs1805007 ^b	338	0,009	0,009	C (0,996)	T (0,004)	1,000
rs1805008 ^b	337	0,045	0,044	C (0,978)	T (0,022)	1,000
rs1805009 ^b	338	0,033	0,032	C (0,984)	G (0,016)	1,000
rs2228479 ^b	337	0,098	0,093	G (0,951)	A (0,049)	1,000
rs11547464 ^b	338	0,006	0,006	G (0,997)	A (0,003)	1,000
rs1110400 ^b	337	0,006	0,006	T (0,997)	C (0,003)	1,000
rs201353883 ^c	338	0,000	0,000	C (1,000)	A (0,000)	-
rs3212355 ^a	337	0,021	0,021	C (0,990)	T (0,010)	1,000
rs3114908 ^d	271	0,358	0,328	C (0,744)	T (0,256)	0,338
rs8051733 ^c	337	0,368	0,363	A (0,763)	G (0,237)	0,880
rs12931267 ^d	336	0,006	0,006	C (0,997)	G (0,003)	1,000
rs2378249 ^b	337	0,193	0,198	A (0,889)	G (0,111)	0,584
rs6119471 ^c	336	0,131	0,196	G (0,890)	C (0,110)	0,000
rs1015362 ^c	337	0,457	0,464	G (0,635)	A (0,365)	0,814
rs6058017 ^a	334	0,296	0,342	A (0,780)	G (0,220)	<u>0,016</u>
rs6059655 ^c	333	0,060	0,064	G (0,967)	A (0,033)	0,301

^a Marcadores genotipados por NGS; ^b Marcadores genotipados por NGS e Array; ^c Marcadores genotipados por Array; ^d Marcadores imputados.

6.4 ANÁLISES FENOTÍPICAS

No contexto das análises preditivas, é importante enfatizar que as curvas ROC são construídas com o intuito de efetuar o cálculo da AUC. Deste modo, todos os valores de AUC serão apresentados, porém, devido ao grande número de curvas ROC construídas, serão apresentadas apenas as curvas correspondentes à maior e menor AUC obtidas. Além disso, os valores de AUC independem de *cut-offs/thresholds* e, portanto, quando estes são aplicados os valores de AUC permanecem os mesmos.

6.4.1 *HirisPlex-S*

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para a ferramenta *HirisPlex-S* estão apresentados na Tabela 8. Em relação à cor dos olhos, três indivíduos não tiveram predição pois não possuíam dados relacionados ao SNP rs12913832, o qual é requisito para a ferramenta realizar a análise. Em relação à cor dos cabelos, três indivíduos não apresentavam dados genotípicos suficientes para a ferramenta realizar a predição. Em relação à cor da pele, indivíduos com valores de p abaixo do *threshold* foram considerados como “não-classificados”. Não foi possível calcular o valor de PPV para a cor de olho intermediária e para a cor de pele muito clara, pois a ferramenta não predisse essas características fenotípicas para nenhum indivíduo.

Combinando os resultados das predições para as três características, isto é, pele, olhos e cabelos, a sensibilidade geral (taxa de acerto) obtida para a ferramenta foi de 0,242.

Tabela 8 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas existentes no sistema *HirisPlex-S*.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Cor dos olhos	Azul	0,914	0,864	0,877	0,404	0,985	0,00%
	Intermediário	0,764	0,000	1,000	-	0,795	0,00%
	Castanho	0,929	0,994	0,630	0,866	0,979	0,00%
	Total	-	0,779	-	-	-	0,00%
Cor dos cabelos	Loiro-Claro	0,882*	0,765	0,925	0,283	0,990	0,00%
	Castanho-Claro	-*	0,244	0,942	0,627	0,757	0,00%
	Castanho-Escuro	0,705*	0,760	0,472	0,470	0,761	0,00%
	Preto	0,834	0,944	0,358	0,352	0,945	0,00%
	Ruivo	0,985	0,667	0,993	0,727	0,991	0,00%
	Total	-	0,660	-	-	-	0,00%
Cor da pele	Muito clara	0,854	0,000	1,000	-	0,873	13,43%
	Clara	0,722	0,138	0,899	0,333	0,740	9,56%
	Intermediária	0,603	0,592	0,500	0,492	0,600	7,21%
	Escura	0,647	0,304	0,818	0,157	0,913	6,12%
	Muito escura	0,984	1,000	0,838	0,263	1,000	0,00%
	Total	-	0,389	-	-	-	8,22%

* O valor da AUC 0,882 refere-se à categoria loiro (claro e escuro), enquanto o valor da AUC 0,705 refere-se à categoria castanho (claro e escuro). Os outros parâmetros são calculados de acordo com as categorias existentes no guia: loiro-claro, castanho-claro (em conjunto com loiro-escuro) e castanho-escuro.

Os resultados obtidos aplicando-se *thresholds* de 0,5 e 0,7 para cor dos olhos estão apresentados na Tabela 9. Combinando as três predições fenotípicas (olhos, cabelos e pele) obtidas aplicando *threshold* de 0,5, tem-se uma taxa de acerto de 0,244. Combinando as três predições obtidas aplicando *threshold* de 0,7, tem-se uma taxa de acerto de 0,269.

Tabela 9 – Sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema *HirisPlex-S*, aplicando *threshold* de 0,5 e 0,7.

Categoria Fenotípica		Sensibilidade	Especificidade	PPV	NPV	Não classificados
Threshold 0,5	Azul	0,902	0,876	0,398	0,990	6,82%
	Intermediário	0,000	1,000	-	0,799	2,94%
	Castanho	0,994	0,650	0,877	0,978	0,00%
	Total	0,787	-	-	-	1,20%
Threshold 0,7	Azul	0,921	0,891	0,443	0,992	13,64%
	Intermediário	0,000	1,000	-	0,832	27,45%
	Castanho	0,994	0,688	0,903	0,975	6,53%
	Total	0,823	-	-	-	11,45%

Em relação aos olhos, a ferramenta se sai muito bem classificando os indivíduos que possuem olhos azuis e castanhos, caracterizados como fenótipos extremos, atingindo uma alta sensibilidade e um alto poder de diferenciação entre essas duas categorias. Como esperado, a ferramenta falha completamente na predição de indivíduos com olhos intermediários, já que

nenhum indivíduo teve este fenótipo predito (Tabela 8). Dentre os 102 indivíduos de olhos verde/mel, 52,94% tiveram predição de olhos azuis (dados desconsiderando a aplicação de *thresholds*), enquanto o restante foi classificado como possuindo olhos castanhos, o que causou uma queda na especificidade dessas duas outras categorias, além de uma queda na sensibilidade geral da ferramenta. O fato das 102 amostras de olhos verde/mel terem sido distribuídas de maneira muito parecida nas categorias “azul” e “castanho” corrobora a ideia de que tais indivíduos possuem variantes tanto relacionadas a cor de olhos azul quanto a cor de olhos castanho, fortalecendo assim o agrupamento desses indivíduos na categoria “intermediário”.

Aplicando os limiares sugeridos de 0,5 e 0,7 pelo artigo original, nota-se que um *threshold* maior aumenta também a sensibilidade total da ferramenta (Tabela 9). No caso do *threshold* 0,7, isso ocorre não pelo fato da sensibilidade dos fenótipos extremos ter aumentado significativamente, mas sim pelo fato de que a maioria dos indivíduos que possuem olhos intermediários havia sido previamente classificados incorretamente com baixas probabilidades de acurácia e deixaram de ser classificados após a aplicação do *threshold*.

Em relação à cor de cabelo, pode-se notar que a ferramenta novamente se sai muito bem ao predizer os fenótipos extremos loiro-claro e preto, atingindo para essas categorias os valores mais altos de sensibilidade. Os valores de PPV para esses fenótipos são mais baixos pois a ferramenta classifica muitas pessoas nessa categoria, o que, de certa forma, contribui para o aumento da sensibilidade. O sistema *HIrisPlex-S* apresenta uma performance insatisfatória na predição dos fenótipos intermediários (castanho-claro e castanho-escuro). Apesar da baixa sensibilidade obtida para cabelos ruivos, a elevada AUC obtida para essa categoria pode ser explicada pela alta especificidade obtida. A sensibilidade geral da ferramenta também foi insatisfatória devido à falha na predição de categorias intermediárias (Tabela 8).

Em relação à cor de pele, o sistema *HIrisPlex-S* apresentou má performance para predição de todas as categorias, falhando completamente na predição de pele muito clara. Para as categorias de pele clara, intermediária e escura, o *HIrisPlex-S* também não apresentou resultados satisfatórios, pois, além de apresentar baixa sensibilidade, também apresentou baixo PPV, ou seja, classificou erroneamente muitas pessoas na categoria em questão. A ferramenta se saiu melhor na predição de pele muito escura. Apesar de ter atingido sensibilidade de 1, é importante lembrar que o baixo valor de PPV indica que a mesma também classificou muitas pessoas com essa tonalidade, o que acaba elevando a sensibilidade.

De todas as ferramentas testadas, a AUC para cor de cabelo ruivo por meio do sistema *HIrisPlex-S* foi a maior obtida. A Figura 5 ilustra a respectiva curva ROC.

Figura 5 – Curva ROC construída para a cor de cor de cabelo ruivo utilizando o sistema *HIrisPlex-S* (AUC = 0,985).

6.4.2 *Snipper6*, *Snipper7*, *Snipper13* e *Snipper23*

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para o conjunto *Snipper6* estão apresentados na Tabela 10. Quando o *cut-off* de 3:1 foi levado em consideração, os indivíduos cujas predições tiveram valores de p abaixo de tal *cut-off* foram considerados como “não-classificados”.

Tabela 10 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do conjunto *Snipper6*, considerando ou não o *cut-off* 3:1.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano Sem cut-off	Azul	0,911	0,864	0,869	0,388	0,985	0,00%
	Intermediário	0,639	0,320	0,756	0,254	0,811	0,00%
	Castanho	0,905	0,720	0,878	0,934	0,566	0,00%
	Total	-	0,651	-	-	-	0,00%
Bayesiano Com cut-off	Azul	0,911	0,964	0,891	0,474	0,996	36,36%
	Intermediário	0,639	0,238	0,882	0,244	0,878	59,22%
	Castanho	0,905	0,850	0,900	0,966	0,643	33,90%
	Total	-	0,776	-	-	-	39,32%
MLR	Azul	0,879	0,881	0,842	0,374	0,985	4,55%
	Intermediário	0,458	0,270	0,746	0,214	0,799	13,59%
	Castanho	0,881	0,270	0,379	0,100	0,668	14,12%
	Total	-	0,630	-	-	-	13,17%

Comparando os parâmetros obtidos pelas duas metodologias preditivas, nota-se que, de maneira geral, o classificador Bayesiano forneceu resultados preditivos melhores, ainda que insatisfatórios. Além disso, ao contrário do modelo preditivo do *IrisPlex*, o *Snipper6* conseguiu prever indivíduos com olhos intermediários, ainda que com um nível de acerto baixo.

Os desenvolvedores dos sistemas preditivos *Snipper* alegam que o uso de um *cut-off* 3:1 poderia melhorar a sensibilidade da ferramenta (Ruiz et al. 2013). Ao aplicarmos esse *cut-off*, a sensibilidade dos olhos azuis e castanhos, assim como a sensibilidade total, aumenta muito, enquanto a dos olhos intermediários decai.

Já o sistema *Snipper7*, como já citado anteriormente, é composto pelo SNP rs1129038 em adição aos seis SNPs do sistema *IrisPlex*. Os desenvolvedores alegam que o haplótipo composto por esse marcador adicional e pelo marcador rs12913832 traria uma melhora significativa na predição dos olhos intermediários (Ruiz et al. 2013).

Ao considerarmos o haplótipo rs1129038-12913832 para predição de cor dos olhos, três amostras foram excluídas das análises envolvendo os sistemas *Snipper7*, *Snipper13* e *Snipper23*, pois seus haplótipos não foram determinados, resultando assim em um espaço amostral de 498 indivíduos.

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para o sistema *Snipper7* estão apresentados na Tabela 11.

Tabela 11 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema *Snipper7*, considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o *cut-off* 3:1.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano – Sem haplótipo e sem cut-off	Azul	0,913	0,841	0,871	0,385	0,983	0,00%
	Intermediário	0,587	0,379	0,673	0,231	0,807	0,00%
	Castanho	0,908	0,641	0,939	0,962	0,521	0,00%
	Total	-	0,605	-	-	-	0,00%
Bayesiano – Sem haplótipo e com cut-off	Azul	0,913	0,902	0,854	0,385	0,902	6,82%
	Intermediário	0,587	0,326	0,725	0,244	0,326	7,77%
	Castanho	0,908	0,688	0,934	0,959	0,688	12,99%
	Total	-	0,631	-	-	-	11,38%
MLR – Sem haplótipo	Azul	0,856	0,860	0,864	0,389	0,984	2,27%
	Intermediário	0,374	0,354	0,672	0,217	0,803	6,80%
	Castanho	0,893	0,636	0,942	0,963	0,520	6,21%
	Total	-	0,599	-	-	-	5,99%
Bayesiano – Com haplótipo e sem cut-off	Azul	0,914	0,841	0,870	0,385	0,983	0,00%
	Intermediário	0,615	0,373	0,646	0,213	0,800	0,00%
	Castanho	0,910	0,611	0,938	0,960	0,500	0,00%
	Total	-	0,582	-	-	-	0,00%
Bayesiano – Com haplótipo e com cut-off	Azul	0,914	0,860	0,864	0,385	0,984	2,27%
	Intermediário	0,619	0,376	0,675	0,238	0,801	0,97%
	Castanho	0,910	0,639	0,944	0,964	0,531	5,62%
	Total	-	0,603	-	-	-	4,42%
MLR – Com haplótipo	Azul	0,856	0,860	0,864	0,389	0,984	2,27%
	Intermediário	0,373	0,354	0,669	0,215	0,802	5,88%
	Castanho	0,894	0,633	0,942	0,963	0,518	5,68%
	Total	-	0,597	-	-	-	5,42%

De todas as ferramentas testadas, a AUC para cor de olho intermediária utilizando o sistema *Snipper7* e MLR como classificadora, considerando haplótipo, foi a mais baixa obtida. A Figura 6 ilustra a respectiva curva ROC.

Figura 6 – Curva ROC construída para a cor de olho intermediária utilizando o sistema *Snipper7*, MLR como classificadora e sem considerar haplótipo (AUC = 0,373).

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para o sistema *Snipper13* estão apresentados na Tabela 12.

Tabela 12 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema *Snipper13*, considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o *cut-off* 3:1.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano – Sem haplótipo e sem <i>cut-off</i>	Azul	0,898	0,636	0,908	0,400	0,963	0,00%
	Intermediário	0,685	0,466	0,709	0,293	0,837	0,00%
	Castanho	0,904	0,701	0,871	0,929	0,547	0,00%
	Total	-	0,647	-	-	-	0,00%
Bayesiano – Sem haplótipo e com <i>cut-off</i>	Azul	0,898	0,762	0,932	0,390	0,986	52,27%
	Intermediário	0,685	0,429	0,752	0,250	0,872	38,83%
	Castanho	0,904	0,741	0,833	0,942	0,470	13,84%
	Total	-	0,692	-	-	-	22,36%
MLR – Sem haplótipo	Azul	0,868	0,881	0,866	0,394	0,987	4,55%
	Intermediário	0,499	0,273	0,813	0,281	0,806	3,88%
	Castanho	0,829	0,791	0,865	0,931	0,642	7,91%
	Total	-	0,690	-	-	-	6,79%
Bayesiano – Com haplótipo e sem <i>cut-off</i>	Azul	0,880	0,841	0,870	0,385	0,983	0,00%
	Intermediário	0,668	0,324	0,712	0,224	0,803	0,00%
	Castanho	0,911	0,679	0,890	0,937	0,535	0,00%
	Total	-	0,620	-	-	-	0,00%
Bayesiano – Com haplótipo e com <i>cut-off</i>	Azul	0,880	0,878	0,863	0,391	0,986	6,82%
	Intermediário	0,668	0,311	0,736	0,228	0,810	11,76%
	Castanho	0,911	0,702	0,916	0,953	0,558	9,38%
	Total	-	0,640	-	-	-	9,64%
MLR – Com haplótipo	Azul	0,869	0,881	0,856	0,398	0,985	4,55%
	Intermediário	0,542	0,304	0,797	0,289	0,809	9,80%
	Castanho	0,831	0,768	0,903	0,946	0,637	15,34%
	Total	-	0,681	-	-	-	13,25%

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para o sistema *Snipper23* estão apresentados na Tabela 13.

Tabela 13 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos olhos, a partir do sistema *Snipper23*, considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o *cut-off* 3:1.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano – Sem haplótipo e sem cut-off	Azul	0,901	0,818	0,882	0,400	0,981	0,00%
	Intermediário	0,690	0,340	0,774	0,280	0,819	0,00%
	Castanho	0,897	0,746	0,850	0,923	0,581	0,00%
	Total	-	0,669	-	-	-	0,00%
Bayesiano – Sem haplótipo e com cut-off	Azul	0,901	0,853	0,908	0,446	0,986	22,73%
	Intermediário	0,690	0,333	0,797	0,260	0,848	27,18%
	Castanho	0,897	0,775	0,826	0,928	0,559	10,73%
	Total	-	0,704	-	-	-	15,17%
MLR – Sem haplótipo	Azul	0,849	0,861	0,865	0,378	0,985	18,18%
	Intermediário	0,502	0,352	0,674	0,226	0,793	14,56%
	Castanho	0,766	0,630	0,903	0,938	0,511	18,36%
	Total	-	0,591	-	-	-	17,56%
Bayesiano – Com haplótipo e sem cut-off	Azul	0,892	0,864	0,870	0,392	0,985	0,00%
	Intermediário	0,663	0,284	0,755	0,230	0,804	0,00%
	Castanho	0,894	0,724	0,863	0,927	0,565	0,00%
	Total	-	0,647	-	-	-	0,00%
Bayesiano – Com haplótipo e com cut-off	Azul	0,892	0,902	0,862	0,394	0,989	6,82%
	Intermediário	0,663	0,277	0,787	0,252	0,807	7,84%
	Castanho	0,894	0,756	0,881	0,938	0,604	9,09%
	Total	-	0,670	-	-	-	8,63%
MLR – Com haplótipo	Azul	0,855	0,889	0,858	0,372	0,988	18,18%
	Intermediário	0,494	0,356	0,676	0,232	0,792	11,76%
	Castanho	0,770	0,629	0,921	0,948	0,518	17,33%
	Total	-	0,592	-	-	-	16,27%

Observando os resultados obtidos para os três sistemas originais *Snipper*, podemos inferir algumas informações gerais: a) o classificador Bayesiano apresentou sensibilidade geral superior à obtida por MLR na maioria das vezes; b) as ferramentas conseguem prever olhos intermediários, ainda que com baixos níveis de sensibilidade; c) ao contrário do esperado, considerar o haplótipo rs1129038-12913832 não traz melhora significativa na predição de olhos verde/mel, além de ter causado uma diminuição na sensibilidade geral das ferramentas na maioria das vezes e d) o uso de *cut-offs* traz uma pequena melhora na sensibilidade geral, embora leve, na maioria das vezes, à redução significativa no número de indivíduos classificados.

Dentre todos os conjuntos, o *Snipper23* foi o que apresentou a maior sensibilidade geral (0,704), aplicando *cut-off* e desconsiderando haplótipo. É interessante observar que o conjunto

aqui considerado como *Snipper6* apresenta sensibilidade maior (0,776) do que qualquer outro conjunto *Snipper* oficialmente proposto.

6.4.3 *Snipper12*

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para o sistema *Snipper12* estão apresentados na Tabela 14. Quando o *cut-off* de 3:1 foi levado em consideração, os indivíduos cujas predições tiveram valores de p abaixo de tal *cut-off* foram considerados como “não-classificados”. Ao considerarmos o haplótipo rs1129038-12913832 para a predição de cor de cabelos, três amostras foram excluídas do conjunto de 462 indivíduos pois seus haplótipos não foram determinados, resultando assim em um espaço amostral de 459 indivíduos.

Tabela 14 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor dos cabelos, a partir do sistema *Snipper12*, considerando ou não o haplótipo rs1129038-12913832 e considerando ou não o *cut-off* 3:1.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano – Sem haplótipo e sem cut-off	Loiro	0,857	0,571	0,890	0,450	0,929	0,00%
	Castanho	0,654	0,481	0,710	0,685	0,511	0,00%
	Preto	0,792	0,752	0,697	0,480	0,883	0,00%
	Ruivo	0,887	0,167	1,000	1,000	0,978	0,00%
	Total	-	0,558	-	-	-	0,00%
Bayesiano – Sem haplótipo e com cut-off	Loiro	0,857	0,673	0,909	0,593	0,934	17,46%
	Castanho	0,654	0,481	0,750	0,664	0,585	38,93%
	Preto	0,792	0,804	0,717	0,557	0,892	22,40%
	Ruivo	0,887	0,143	1,000	1,000	0,981	41,67%
	Total	-	0,604	-	-	-	31,60%
MLR – Sem haplótipo	Loiro	0,604	0,161	0,986	0,692	0,853	11,11%
	Castanho	0,601	0,576	0,648	0,632	0,593	35,11%
	Preto	0,801	0,895	0,713	0,556	0,944	24,00%
	Ruivo	0,778	0,455	0,981	0,455	0,981	8,33%
	Total	-	0,604	-	-	-	29,44%
Bayesiano – Com haplótipo e sem cut-off	Loiro	0,854	0,635	0,886	0,471	0,938	0,00%
	Castanho	0,674	0,488	0,773	0,738	0,535	0,00%
	Preto	0,806	0,798	0,701	0,497	0,903	0,00%
	Ruivo	0,845	0,182	1,000	1,000	0,980	8,33%
	Total	-	0,585	-	-	-	0,22%
Bayesiano – Com haplótipo e com cut-off	Loiro	0,854	0,667	0,876	0,486	0,937	14,29%
	Castanho	0,674	0,466	0,784	0,714	0,560	25,77%
	Preto	0,806	0,832	0,723	0,560	0,910	13,71%
	Ruivo	0,845	0,167	1,000	1,000	0,986	50,00%
	Total	-	0,600	-	-	-	21,57%
MLR – Com haplótipo	Loiro	0,619	0,170	0,985	0,692	0,854	15,87%
	Castanho	0,599	0,535	0,654	0,612	0,580	38,85%
	Preto	0,804	0,905	0,682	0,551	0,943	23,39%
	Ruivo	0,780	0,500	0,990	0,571	0,987	33,33%
	Total	-	0,584	-	-	-	31,37%

Observando a Tabela 14, nota-se que a introdução de haplótipo traz pioras quando se usa o classificador MLR ou quando se aplica um *cut-off* aos resultados Bayesianos. Porém, o haplótipo traz um leve aumento na sensibilidade geral dos resultados Bayesianos sem a aplicação de *cut-offs*. A maior sensibilidade geral foi atingida ao usarmos o classificador Bayesiano, sem a introdução de haplótipos e aplicando o *cut-off* de 3:1.

6.4.4 *Snipper10*

Para estratificação das amostras, seguindo as recomendações dos desenvolvedores da ferramenta, aquelas com valor de L^*_{ab} superior a 60,00 foram classificadas como possuindo pele

branca; um valor de L^*_{ab} entre 40,00 e 60,00 foi classificado como pele intermediária; um valor de L^*_{ab} abaixo de 40,00 foi considerado pele escura.

Os parâmetros de AUC, sensibilidade, especificidade, PPV e NPV obtidos para o sistema *Snipper10* estão apresentados na Tabela 15. Quando o *cut-off* de 3:1 foi levado em consideração, os indivíduos cujas predições tiveram valores de p abaixo de tal *cut-off* foram considerados como “não-classificados”.

Tabela 15 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor de pele, a partir do sistema *Snipper10*, considerando ou não o *cut-off* 3:1.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano – Sem <i>cut-off</i>	Branca	0,849	0,908	0,582	0,838	0,726	0,00%
	Intermediária	0,842	0,507	0,907	0,683	0,823	0,00%
	Escura	0,773	0,333	0,977	0,154	0,992	0,00%
	Total	-	0,787	-	-	-	0,00%
Bayesiano – com <i>cut-off</i>	Branca	0,849	0,919	0,567	0,855	0,716	3,75%
	Intermediária	0,842	0,521	0,920	0,693	0,847	16,43%
	Escura	0,773	0,000	0,984	0,000	0,993	50,00%
	Total	-	0,811	-	-	-	7,91%
MLR	Branca	0,742	0,914	0,481	0,831	0,667	16,14%
	Intermediária	0,643	0,198	0,956	0,606	0,776	27,86%
	Escura	0,758	0,667	0,270	0,007	0,991	50,00%
	Total	-	0,729	-	-	-	19,88%

O classificador Bayesiano apresentou uma melhor performance do que o MLR, apresentando maior sensibilidade geral. Como esta análise não envolve a possibilidade de haplótipos compostos pelos marcadores rs12913832 e rs1129038, os quais não fazem parte do sistema, a única variável a se analisar é o estabelecimento de um *cut-off* 3:1. Ao ser utilizado, a sensibilidade geral aumenta devido ao aumento de sensibilidade para peles brancas e intermediárias, porém a sensibilidade para pele escura é zero. A ferramenta prediz sete indivíduos com pele escura, porém nenhum deles apresenta o fenótipo em questão.

Além da classificação dos indivíduos de acordo com o valor L^*_{ab} , os indivíduos também foram estratificados seguindo a classificação de Fitzpatrick (utilizada pelo sistema *HirisPlex-S*), para verificar se a performance da ferramenta seria muito diferente. Os resultados estão apresentados na Tabela 16.

Tabela 16 – AUC, sensibilidade, especificidade, PPV e NPV obtidos para as categorias fenotípicas correspondentes a cor de pele, a partir do sistema *Snipper10*, considerando ou não o *cut-off* 3:1 e estratificando os indivíduos de acordo com a classificação de Fitzpatrick.

Categoria Fenotípica		AUC	Sensibilidade	Especificidade	PPV	NPV	Não classificados
Bayesiano – Sem <i>cut-off</i>	Branca	0,843	0,975	0,392	0,524	0,959	0,00%
	Intermediária	0,641	0,248	0,809	0,509	0,573	0,00%
	Escura	0,885	0,162	0,998	0,923	0,872	0,00%
	Total	-	0,531	-	-	-	0,00%
Bayesiano – com <i>cut-off</i>	Branca	0,843	0,990	0,369	0,546	0,980	1,97%
	Intermediária	0,641	0,227	0,824	0,495	0,584	10,81%
	Escura	0,885	0,113	1,000	1,000	0,878	16,22%
	Total	-	0,542	-	-	-	8,02%
MLR	Branca	0,697	0,960	0,314	0,526	0,909	12,81%
	Intermediária	0,531	0,098	0,920	0,486	0,570	21,62%
	Escura	0,688	0,449	0,943	0,524	0,925	33,78%
	Total	-	0,523	-	-	-	19,84%

Pode-se observar que utilizando essa classificação, a sensibilidade geral atingida pela ferramenta nos três casos é muito insatisfatória. Apesar da sensibilidade ter aumentado para peles brancas, o sistema continua apresentando dificuldades na predição de peles intermediárias e escuras.

Combinando as predições do *Snipper23*, *Snipper12* e *Snipper10* para as três características, isto é, olhos, cabelos e pele, respectivamente, aplicando o *cut-off* e desconsiderando a fase haplotípica dos SNPs rs12913832 e rs1129038 em todos eles, atinge-se uma sensibilidade (taxa de acerto) de 0,298. Este valor é um pouco maior que o obtido pelo *HirisPlex-S* (0,269), porém envolve a desclassificação de 46,51% dos indivíduos, contra 11,45% pelo *HirisPlex-S*.

6.5 ANÁLISES DE ASSOCIAÇÃO

As análises de associação foram realizadas para as 468 amostras que possuíam dados prévios de ancestralidade. Associações alélicas adicionando ou não a ancestralidade como fator de confusão foram realizadas por meio de uma MLR utilizando o *software* plink v1.07.

A fim de verificar se os marcadores presentes nas ferramentas preditivas realmente estão associados com a pigmentação nesta amostra da população brasileira, ou se na verdade tem-se uma associação enviesada pela ancestralidade, estão mostradas na Tabela 17 as associações alélicas encontradas adicionando a ancestralidade individual como fator de confusão. O

conjunto completo de associações alélicas obtidas sem ajuste por ancestralidade está localizada no Apêndice B.

Além disso, no Apêndice C podem ser encontradas o número de associações obtidas para cada marcador, divididas por ajustadas/não ajustadas por ancestralidade e, quando não ajustadas, divididas também entre associações fracas ($1 < OR < 5$), moderadas ($5 \leq OR < 10$) e fortes ($OR \geq 10$).

Tabela 17 – Associações alélicas significativas encontradas para os 61 marcadores dos seis conjuntos preditivos analisados, adicionando a ancestralidade como um fator de confusão.

Marcador	Sistema preditivo *	Alelo	Característica	<i>p</i> (teste exato de Fisher)	<i>Odds Ratio</i>	IC 95%
rs2238289	C	G	Olho Intermediário vs. Azul	$1,78 \times 10^{-2}$	6,18	1,37-27,9
rs7183877	E,F	A	Olho Castanho vs. Azul	$1,40 \times 10^{-2}$	12,48	1,66-93,66
rs2238289	C	G	Olho Castanho vs. Azul	$9,10 \times 10^{-6}$	26,34	6,21-111,70
rs12931267	G	G	Cabelo Ruivo vs. Castanho Claro	$5,13 \times 10^{-4}$	75,41	6,58-864,40
rs1805007	B, C, G	T	Cabelo Ruivo vs. Castanho Claro	$3,46 \times 10^{-5}$	412,80	23,86-7140
rs7495174	F, G	G	Cabelo Ruivo vs. Castanho Claro	$1,87 \times 10^{-3}$	5,82	1,918-17,66
rs8051733	C	G	Cabelo Ruivo vs. Castanho Claro	$5,78 \times 10^{-5}$	11,57	3,51-38,14
rs1129038	C, D, E, F, G	T	Cabelo Loiro Claro vs. Ruivo	$1,33 \times 10^{-2}$	26,69	1,98-359,30
rs12913832	A, B, C, D, E, F, G	G	Cabelo Loiro Claro vs. Ruivo	$1,33 \times 10^{-2}$	26,69	1,98-359,30
rs11636232	E, F	T	Cabelo Loiro Claro vs. Ruivo	$3,27 \times 10^{-2}$	9,32	1,20-72,30
rs12896399	A, C, D, E, F	T	Cabelo Loiro Claro vs. Castanho Claro	$2,86 \times 10^{-2}$	2,33	1,09-4,96
rs1129038	C, D, E, F, G	T	Cabelo Castanho Claro vs. Preto	$3,68 \times 10^{-6}$	4,44	2,36-8,35
rs4778232	E, F	T	Cabelo Castanho Escuro vs. Castanho Claro	$1,75 \times 10^{-3}$	1,88	1,27-2,79
rs4778241	E, F	C	Cabelo Castanho Escuro vs. Castanho Claro	$9,13 \times 10^{-3}$	2,56	1,26-5,18
rs7495174	F, G	G	Cabelo Castanho Escuro vs. Castanho Claro	$7,26 \times 10^{-4}$	2,74	1,53-4,92
rs6497292	C	G	Cabelo Castanho Escuro vs. Castanho Claro	$4,89 \times 10^{-4}$	2,79	1,57-4,98
rs916977	F	T	Cabelo Castanho Escuro vs. Castanho Claro	$4,45 \times 10^{-6}$	2,60	1,73-3,91
rs28777	B, C, G	C	Cabelo Preto vs. Castanho Claro	$8,76 \times 10^{-3}$	2,15	1,21-3,82
rs2402130	B, C, H	G	Cabelo Preto vs. Castanho Claro	$1,80 \times 10^{-2}$	2,11	1,14-3,91
rs2238289	C	G	Cabelo Preto vs. Castanho Claro	$6,34 \times 10^{-3}$	2,47	1,29-4,74
rs6497292	C	G	Cabelo Preto vs. Castanho Claro	$2,16 \times 10^{-4}$	4,10	1,94-8,66
rs12592730	F	A	Cabelo Preto vs. Castanho Claro	$4,55 \times 10^{-3}$	4,28	1,57-11,69
rs2378249	B,C	G	Pele Clara vs. Intermediária	$1,92 \times 10^{-2}$	1,97	1,12-3,47
rs1129038	C, D, E, F, G	T	Pele Clara vs. Não clara	$2,83 \times 10^{-4}$	1,88	1,34-2,64
rs8024968	E, F	T	Pele Intermediária vs. Clara	$1,12 \times 10^2$	1,78	1,14-2,79
rs1470608	C	T	Pele Intermediária vs. Clara	$4,52 \times 10^{-2}$	1,50	1,01-2,22

* A = *IrisPlex*; B = *HIrisPlex*; C = *HIrisPlex-S*; D = *Snipper7*; E = *Snipper13*; F = *Snipper23*; G = *Snipper12*; H = *Snipper10*.

6.6 ANÁLISES DE ANCESTRALIDADE

O *software* STRUCTURE foi utilizado para estimar ancestralidade a níveis populacionais e individuais. As análises foram realizadas separadamente para cada uma das seis ferramentas preditivas. Também foi realizada uma análise incluindo todos os 61 marcadores presentes nas ferramentas. As figuras 7, 8 e 9 ilustram os diagramas triangulares e os gráficos em barra representando as estimativas de ancestralidade de cada indivíduo obtidas a partir de cada conjunto de marcadores. A Tabela 18 apresenta os valores de proporções de ancestralidade populacional obtidos a partir dos genótipos dos marcadores que compõem cada ferramenta preditiva.

Figura 7 – Diagramas triangulares gerados pelo *software* STRUCTURE a partir dos seis sistemas de marcadores preditivos analisados, representando as estimativas de ancestralidade geradas para cada indivíduo. Os círculos coloridos representam os indivíduos, divididos em populações por meio das cores: amarelo – indivíduos provenientes de Ribeirão Preto, vermelho – europeus (EUR), azul – asiáticos (EAS), verde – africanos (AFR).

Figura 8 – Diagrama triangular gerado pelo software STRUCTURE a partir dos 61 marcadores preditivos analisados, representando as estimativas de ancestralidade geradas para cada indivíduo.

Figura 9 – Gráficos retangulares gerados pelo *software* distruct 1.1, baseados nas análises realizadas pelo *software* STRUCTURE a partir dos seis sistemas de marcadores preditivos analisados, representando as estimativas de ancestralidade geradas para cada indivíduo. As barras verticais representam cada indivíduo e assumem até três cores, de acordo com a ancestralidade do mesmo: vermelho, verde e azul correspondem a ancestralidade europeia (EUR), africana (AFR) e asiática (EAS), respectivamente.

Tabela 18 – Estimativas de ancestralidade para a amostra populacional de Ribeirão Preto – SP, inferidas pelo *software* STRUCTURE, utilizando os marcadores presentes nas seis ferramentas preditivas analisadas.

Sistema preditivo	Europeu	Africano	Asiático
<i>Snipper7</i>	0,539	0,287	0,174
<i>Snipper13</i>	0,443	0,237	0,320
<i>Snipper23</i>	0,512	0,238	0,250
<i>Snipper10</i>	0,705	0,158	0,137
<i>Snipper12</i>	0,496	0,368	0,135
<i>HIrisPlex-S</i>	0,719	(0,281)*	(0,281)*
61 marcadores	0,619	0,211	0,171

* Como o resultado da análise não permite separar as populações africana e asiática, o valor de 0,281 representa a contribuição das duas populações juntas.

Discussão

7. Discussão

7.1 IMPUTAÇÃO

Comparando os dois programas utilizados para imputar os genótipos dos marcadores não genotipados diretamente, o IMPUTE2 apresentou maior facilidade de uso, pois, apesar de consistir em várias etapas, os resultados são obtidos rapidamente (levando menos de um dia) e todo o procedimento é realizado por meio de scripts, reduzindo o trabalho manual da preparação dos arquivos para a imputação. Devido a isso, é possível imputar regiões maiores com maior facilidade e sem haver diferença significativa entre os tempos de análise. Porém, esse *software* exige muita memória RAM para ser executado. O *Michigan Imputation Server* (MIS) gera resultados ainda mais rapidamente que a outra ferramenta, além de constituir plataforma online, podendo ser acessado de qualquer computador e não ocupando os recursos computacionais disponíveis no laboratório. Porém, apesar de ser de fácil preparo, o arquivo de entrada para esse programa deve ser conferido manualmente para que se garanta que os marcadores apresentem os mesmos alelos referência/alternativo na população a ser analisada e nas populações de referência. Enquanto o MIS gera resultados para todos os marcadores alvo em todos os indivíduos, o IMPUTE2 resulta em genótipos interrogados para alguns indivíduos. A princípio, o IMPUTE2 apresentou índices de qualidade teórica maiores para os marcadores alvo, mas ainda assim próximos dos resultados obtidos com o MIS.

Considerando os resultados obtidos para os marcadores usados para testar e validar o processo de imputação (Tabela 3), o IMPUTE2 apresenta maiores índices de qualidade teórica em todos os casos. Porém, ao avaliar a acurácia real, enquanto o MIS apresentou acurácia média real maior que a teórica, o IMPUTE2 apresentou acurácia real menor que a qualidade teórica. Apesar disso, os valores obtidos ainda são excelentes. Ao unificar os genótipos obtidos pelos dois métodos, ou seja, aceitar considerar como sendo verdadeiros apenas os genótipos que foram imputados igualmente pelas duas ferramentas, a acurácia obtida aumenta para 0,990, com apenas 2,60% de genótipos interrogados.

Considerando os 11 marcadores cujos genótipos precisavam realmente ser imputados, os métodos MIS e IMPUTE2 apresentam pequena taxa média de discordância (2,41%), variando de 0,20% a 8,70% (Tabela 5). Em apenas dois casos a porcentagem de discordâncias é grande: rs4778241 (8,70%) e rs3114908 (6,22%). Esses são dois marcadores que mostram

uma menor qualidade teórica de imputação (0,809 e 0,920 respectivamente) com o IMPUTE2 e em ambos os casos essa ferramenta gerou um grande número de genótipos interrogados. Entretanto, o ensaio de validação não mostrou relação da acurácia com a qualidade teórica ou proporção de genótipos interrogados. A fim de minimizar erros associados ao processo de imputação, ao invés de utilizar os resultados obtidos por apenas um dos métodos, foram utilizados os resultados concordantes entre as duas metodologias.

7.2 DESVIOS DO EQUILÍBRIO DE HARDY-WEINBERG

Dos 61 sítios de variação analisados, 8 deles (13,11%) apresentaram desvios em relação ao equilíbrio de Hardy-Weinberg: rs16891982 (*SLC45A2*), rs28777 (*SLC45A2*), rs3829241 (*TPCN2*), rs1426654 (*SLC24A5*), rs1800414 (*OCA2*), rs4778241 (*OCA2*), rs6119471 (*ASIP*) e rs6058017 (*ASIP*) (Tabela 7). Dos oito *loci*, seis apresentam deficiência de heterozigotos. O marcador rs1800414 (T: 0,9905 e C: 0,0095) apresenta um alelo raro em nossa população. Apesar de que seria esperado que tal alelo raro fosse observado apenas em heterozigose, um homozigoto (além de três heterozigotos) para o alelo raro foi identificado. Quando isso ocorre, a deficiência pode estar sendo causada por um erro de genotipagem (heterozigotos erroneamente genotipados como homozigotos) (Graffelman et al. 2015). Entretanto, é importante ressaltar que o marcador rs1800414 foi genotipado por duas metodologias independentes (NGS e MEGA Array), o que elimina esta possibilidade de erro. Os outros marcadores, por outro lado, possuem alelos com frequências melhor distribuídas. É importante ressaltar também que os oito marcadores em desequilíbrio foram genotipados por metodologias diferentes: dois por NGS, três por MEGA Array, dois por NGS e MEGA Array e um por imputação. Essa distribuição nos leva a afirmar que o desequilíbrio observado aqui não possui uma causa metodológica, ou seja, não é causada por erro de genotipagem. Além disso, ao ser aplicada a correção de Bonferroni para os 61 marcadores ($\alpha_c = 0,05/61 = 0,00082$), seis dos oito marcadores que apresentavam desvios em relação ao equilíbrio deixaram de apresentá-lo.

Por fim, o fato de todos esses marcadores fazerem parte de conjuntos específicos de predição da cor de olho, cabelo e/ou pele sugere que os mesmos devem possuir fortes associações com algumas dessas características, o que pode indicar algum efeito de seleção natural em marcadores específicos. De fato, dois desses marcadores (rs16891982 e rs1426654)

são reconhecidos como marcadores muito relacionados à pigmentação (Graf et al. 2005; Norton et al. 2006; Cook et al. 2009; Spichenok et al. 2011; Walsh, Liu, et al. 2011). Esta hipótese ainda é fortalecida pelo fato de outros marcadores também estarem presentes em alguns dos principais genes reconhecidos por influenciar amplamente na pigmentação: *SLC45A2*, *OCA2* e *ASIP* (Parra 2007; Sturm 2009; Liu et al. 2013). A única exceção seria o SNP rs3829241, pertencente ao gene *TPCN2*. Esse gene, apesar de não estar entre os genes mais reconhecidos ligados a pigmentação, codifica a proteína TPC2, e existem estudos que mostram que essa proteína influencia na via de pigmentação, regulando o pH e o tamanho dos melanossomos (Ambrosio et al. 2016; Bellono et al. 2016). Além disso, também existem estudos na literatura que mostram a associação deste SNP com a pigmentação (Sulem et al. 2008; Branicki et al. 2011).

7.3 PREDIÇÕES FENOTÍPICAS

Primeiramente, é importante ressaltar que os marcadores que compõem os sistemas de predição foram selecionados por meio de estudos realizados em populações europeias (Walsh, Liu, et al. 2011; Ruiz et al. 2013; Walsh et al. 2013; Maroñas et al. 2014; Söchtig et al. 2015; Chaitanya et al. 2018). As únicas ferramentas que utilizaram também amostras de populações não-europeias para desenvolver seus modelos estatísticos foram *Snipper10* e *HirisPlex-S* (apenas para predição da cor da pele). Como os outros conjuntos de marcadores foram desenvolvidos exclusivamente com base em populações europeias (Walsh, Liu, et al. 2011; Ruiz et al. 2013; Walsh et al. 2013; Söchtig et al. 2015), é esperado que as ferramentas apresentem desempenho satisfatório quando usadas na Europa, porém os resultados podem ser bem diferentes quando testadas em outras populações.

Nos estudos envolvendo as ferramentas *Snipper* estão disponíveis valores de sensibilidade, os quais serão discutidos posteriormente. Em relação ao sistema *HirisPlex-S*, os dados mais atuais e concretos disponíveis para comparação com os resultados aqui obtidos são os valores de AUC, disponíveis no site da ferramenta online (<https://hirisplex.erasmusmc.nl/>) (Tabela 19).

Tabela 19 – AUCs estimados pelos desenvolvedores do sistema *HirisPlex-S* e sensibilidades obtidas para os sistemas *Snipper* para as diferentes categorias de cada característica fenotípica em comparação com os valores obtidos no presente estudo.

Característica fenotípica	AUC – <i>HirisPlex-S</i>		Sensibilidade – <i>Snipper</i>	
	Desenvolvedores	Presente estudo	Desenvolvedores	Presente estudo
Olho azul	0,94	0,91	1,00/1,00/1,00*	0,86/0,88/0,90*
Olho intermediário	0,74	0,76	0,32/0,26/0,21*	0,38/0,31/0,28*
Olho castanho	0,95	0,93	0,78/0,82/0,80*	0,64/0,70/0,76*
Cabelo loiro	0,81	0,88	0,85	0,67
Cabelo castanho	0,74	0,70	0,45	0,47
Cabelo ruivo	0,93	0,98	0,78	0,17
Cabelo preto	0,86	0,83	0,75	0,83
Pele muito clara	0,83	0,85	-	-
Pele clara	0,76	0,72	0,97	0,92
Pele intermediária	0,78	0,60	0,62	0,20
Pele escura	0,98	0,65	0,30	0,67
Pele muito escura	0,99	0,98	-	-
Média	0,86	0,82	0,68	0,58

OBS: As sensibilidades marcadas com * representam valores para *Snipper7*, *Snipper13* e *Snipper23*, respectivamente.

7.3.1 – Cor dos olhos

Em relação à cor dos olhos, os valores de AUC obtidos para as cores azul e castanho (Tabela 19) se mostraram um pouco inferiores (aproximadamente 0,02 unidades) aos valores estimados pelos desenvolvedores, enquanto a AUC para olhos intermediários foi levemente superior, apesar de ainda não ser satisfatória. Os artigos originais da ferramenta (Walsh, Liu, et al. 2011; Walsh, Lindenberg, et al. 2011) já haviam citado dificuldades na predição da cor intermediária (verde/mel), e esse fato pode ser confirmado neste trabalho: o *HirisPlex-S* falhou completamente na predição de olhos intermediários: nenhum indivíduo no geral foi classificado com este fenótipo (sensibilidade 0 e especificidade 1), o que indica que, neste caso, o valor de AUC é um tanto quanto ilusório. Apesar de não corresponder a um valor muito alto (0,76), à primeira vista pode-se ter a impressão de que a ferramenta consegue ter uma performance ao menos razoável na predição de olhos intermediários, porém ela falha completamente. Isto porque o gráfico ROC não reflete imediatamente a acurácia da predição, mas sim serve para indicar qual seria o melhor *cut-off* de uso para uma ferramenta, considerando que a mesma é

um classificador binário (comparando apenas “cor x” contra “cor não x”). Como o *HIrisPlex-S* não consiste em um classificador binário, possuindo duas outras categorias para cor dos olhos, mesmo que o gráfico ROC indicasse que um *cut-off* de $p = 0,3$, por exemplo, para olhos intermediários seria o ideal para classificar um indivíduo como possuindo ou não o fenótipo em questão, isto não seria facilmente aplicável, pois os *cut-offs* das outras duas categorias teriam de ser levados em consideração.

O fato do sistema possuir grande dificuldade na predição de olhos intermediários também é corroborado por estudos feitos em outras populações, sejam elas europeias (Itália, Portugal e Eslovênia) ou não (Ásia, Brasil, EUA e Venezuela) (Kastelic et al. 2013; Dembinski e Picard 2014; Freire-Aradas et al. 2014; Yun et al. 2014; Dario et al. 2015; Salvoro et al. 2019), e é especialmente importante quando se trata de populações miscigenadas, pois esta característica faz com que fenótipos intermediários estejam muito presentes nessas populações. Olhando para a presente amostragem, pode-se notar que o número de indivíduos coletados com olhos azuis (44) corresponde a menos da metade de indivíduos coletados com olhos verde/mel (103). Esses achados nos levam à hipótese de que no sistema *HIrisPlex-S* faltam marcadores que estejam fortemente associados com as cores verde/mel e que sejam capazes de diferenciá-la das outras duas categorias.

Um trabalho realizado em 2014 (Freire-Aradas et al. 2014) analisou o desempenho do *IrisPlex* em amostras populacionais brasileira ($n = 79$) e venezuelana ($n = 20$), considerando apenas as predições cujas probabilidades ultrapassassem um *threshold* de 0,7, a fim de aumentar a especificidade. Nele, a ferramenta atinge uma sensibilidade maior do que a obtida em nosso estudo para olhos azuis (1,000) com especificidade menor (0,718), enquanto o inverso ocorre para a cor castanha (menor sensibilidade e maior especificidade do que o observado no presente estudo). Novamente, o *IrisPlex* falhou completamente na predição de olhos intermediários.

Em relação aos conjuntos *Snipper* testados para predição da cor dos olhos, o conjunto *Snipper6* (classificador Bayesiano, *cut-off* 3:1) se mostrou melhor que os outros conjuntos, atingindo a maior sensibilidade total, porém não classificando 39,32% das amostras. Comparando seus resultados com aqueles obtidos pelo *IrisPlex*, este último se mostra uma melhor ferramenta preditiva pois apresenta valores de AUC e sensibilidade total superiores àqueles apresentados pelo *Snipper6*. As duas ferramentas apresentam performance semelhante na determinação de olhos azuis sem aplicação de *cut-off*, porém o *Snipper* atinge uma acurácia muito superior quando o *cut-off* é levado em consideração. Enquanto essa mesma ferramenta apresenta a capacidade de predição de olhos intermediários (ao contrário do *IrisPlex*) ainda que

com baixa AUC e sensibilidade, o sistema *IrisPlex* apresenta uma performance muito superior na determinação de olhos castanhos. Apesar dos dois conjuntos possuírem os mesmos seis marcadores, o fato dos modelos matemáticos usados não serem os mesmos (classificador Bayesiano e MLR com diferentes ajustes) explica a diferença entre os resultados.

No trabalho de Freire-Aradas et al. (2014) citado anteriormente, o *Snipper6* também foi testado (considerando o classificador Bayesiano com *cut-off* 3:1) nas amostras populacionais brasileira e venezuelana, apresentando maior sensibilidade apenas para a cor azul (apesar de apresentar AUC inferior) quando comparado ao nosso estudo. Apesar da sensibilidade para olhos intermediários ser inferior à obtida aqui (0,125 contra 0,238), a AUC é maior (0,729 contra 0,639).

O *Snipper7* também já havia sido testado em população sul americana, utilizando o classificador Bayesiano e considerando o *cut-off* 3:1 e haplótipo composto pelos marcadores rs1129038-12913832. No estudo realizado por Freire-Aradas em 2014, a ferramenta apresentou melhor performance na predição de maneira geral, apresentando apenas um valor de sensibilidade para olhos intermediários mais baixo, acompanhado também de um valor de AUC insatisfatório (0,628).

Como os conjuntos *Snipper13* e *Snipper23* consistem em uma versão do *Snipper7* com mais marcadores, seria esperado que as predições melhorassem conforme o conjunto de SNPs aumentasse. Entretanto, apesar de valores totais de sensibilidade terem se tornado maiores em alguns casos, esse aumento não é expressivo. O fato do aumento do número de marcadores não trazer uma melhora expressiva na performance da ferramenta é corroborado também pelo estudo feito por Freire-Aradas *et al.* (2014).

O diferencial esperado para os conjuntos *Snipper* seria a capacidade de predizer olhos intermediários. Porém, dentre os conjuntos *Snipper13* e *Snipper23*, a maior sensibilidade obtida para essa característica é de 0,466 (Tabela 12). Os baixos valores de AUC que acompanham os níveis insatisfatórios de sensibilidade obtidos para os olhos intermediários indicam que essa predição está mais próxima de uma classificação aleatória (o que é indicado por $AUC = 0,5$).

De maneira geral, os três conjuntos originais *Snipper* para predição de cor dos olhos aqui analisados apresentaram uma performance inferior àquela apresentada pelo *HIrisPlex-S*. O melhor conjunto apresentou sensibilidade total de 0,704 (*Snipper23*, com *cut-off*, sem considerar haplótipo). Apesar de serem eficientes na determinação dos olhos azuis, os resultados da determinação de olhos castanhos não foram satisfatórios e a capacidade de

predição de olhos intermediários, que seria o grande diferencial das ferramentas, mostrou baixos índices de sensibilidade acompanhados de valores de AUC próximos a 0,5. Porém, em todas as possibilidades analisadas, mesmo conseguindo prever alguns indivíduos com olhos intermediários corretamente, as ferramentas apresentaram valores de sensibilidade total muito baixos, inviabilizando seu uso na rotina forense em nossa população.

Apesar do sistema *IrisPlex* ter apresentado muito boa sensibilidade na predição de olhos castanhos (e azuis, quando aplicando *threshold*), sua completa falha na predição de olhos intermediários diminui a especificidade das outras categorias, além da sensibilidade total, que beira os 80%, inviabilizando seu uso em populações miscigenadas.

7.3.2 – Cor dos cabelos

Considerando-se o sistema *HIrisPlex*, os valores de AUC obtidos para as cores loiro e ruivo se mostraram superiores (aproximadamente 0,06 unidades) aos valores estimados pelos desenvolvedores, enquanto que os valores de AUC obtidos para cabelos castanhos e pretos foram inferiores (Tabela 19). Assim como no caso de cor dos olhos, os fenótipos preditos com maior taxa de acerto foram os fenótipos extremos (loiro-claro/castanho-escuro).

No caso do conjunto *Snipper12* para predição da cor do cabelo, as AUCs obtidas para as quatro cores se mostraram inferiores ao serem comparadas com os resultados do *HIrisPlex-S*. A ferramenta demonstrou sensibilidade maior que 0,8 apenas para cor de cabelo preta. Apesar da sensibilidade para cabelos ruivos ser baixa no geral, a especificidade e o PPV assumem valores altos, indicando que quando a ferramenta prediz cabelos ruivos para algum indivíduo, esse indivíduo realmente é ruivo. Pode-se observar que a ferramenta também apresenta dificuldade na predição de cabelos loiro, e principalmente, castanho, com sensibilidades muito baixas (máximo de 0,673 para loiro e 0,576 para castanho). Parte disso talvez seja decorrente da forma de classificação das cores intermediárias, pois enquanto no *HIrisPlex-S* tem-se a divisão de “loiro-claro”, “loiro-escuro/castanho-claro” e “castanho-escuro”, no *Snipper12* as cores são divididas apenas em “loiro” e “castanho”. O *HIrisPlex-S* junta as categorias “loiro-escuro/castanho-claro” pois elas são muito parecidas e difíceis de serem diferenciadas a olho nu (Walsh et al. 2013; Walsh et al. 2014). Inclusive, é para essa categoria intermediária (loiro-escuro/castanho-claro) que o *HIrisPlex-S* apresenta maior dificuldade de predição (sensibilidade de 0,244), apesar de apresentar sensibilidade acima de 0,750 para as categorias

“loiro-claro” e “castanho-escuro”. A dificuldade de diferenciação das cores somada ao fato de elas terem que ser separadas simplesmente em “loiro” e “castanho” no *Snipper* pode estar contribuindo para as baixas sensibilidades apresentadas para essas cores.

Apesar dos valores totais de sensibilidade atingidos serem baixos, no estudo original, que usa o classificador Bayesiano considerando o haplótipo composto pelos marcadores rs1129038-12913832 e o *cut-off* 3:1, a sensibilidade atingida foi de 0,615, com 20,3% dos indivíduos não classificados (contra 0,600 em nossa população com 21,57% de amostras não classificadas). No estudo original, os seguintes valores de AUC foram obtidos: 0,939 para ruivo, 0,850 para loiro, 0,842 para preto e 0,642 para castanho. As AUCs obtidas para ruivo e preto no presente estudo se mostraram inferiores à do estudo original.

Dos 12 SNPs presentes neste conjunto, 7 também fazem parte do sistema *HIrisPlex*. É importante ressaltar que os desenvolvedores do *Snipper12* deixam claro que a intenção da ferramenta não é substituir o *HIrisPlex*, mas sim explorar e contribuir com o melhor entendimento do papel de alguns marcadores adicionais na predição das cores de cabelo (Söchtig et al. 2015).

Infelizmente, apesar de estudos já terem sido feitos analisando a acurácia do sistema *IrisPlex* e dos conjuntos *Snipper* para predição de cor dos olhos em outras populações, nenhum estudo foi realizado com o âmbito de analisar em outras populações os sistemas *HIrisPlex* e *Snipper12* (cor de cabelo), bem como os *HIrisPlex-S* e *Snipper10* (cor de pele), que serão discutidos no item a seguir.

7.3.3 – Cor da pele

Considerando-se o sistema *HIrisPlex-S*, apesar da AUC obtida para pele muito clara ser a única superior ao valor estimado pelos desenvolvedores, nenhum indivíduo foi classificado com este fenótipo (sensibilidade 0 e especificidade 1). Isso pode ser explicado de maneira similar ao caso dos olhos intermediários: de acordo com a curva ROC, o melhor *cut-off* para classificar um indivíduo com pele muito clara, balanceando os valores de sensibilidade e especificidade, seria um *p-value* aproximado de 0,0140, o que garantiria sensibilidade e especificidade de aproximadamente 0,776 e 0,782, respectivamente. Porém, esse valor é muito baixo, e ao aplicarmos a ferramenta, existem outras categorias que possuem *p-values* muito superiores a esse e que serão assumidas como o fenótipo do indivíduo em questão.

O sistema *HIrisPlex-S* apresentou má performance para predição da cor da pele no geral. É importante lembrar que, mesmo que nosso grupo de pesquisa tenha usado o mesmo esquema de classificação que os desenvolvedores do sistema *HIrisPlex-S* (escala Fitzpatrick), fenótipos de pele extremamente claras são mais comuns na Europa do que no Brasil. Nosso país, por outro lado, apesar de possuir uma parcela muito grande de indivíduos brancos, também apresenta muitas pessoas com fenótipos extremamente escuros, o que é pouco frequente na Europa. Por isso, a percepção de uma pele muito clara e uma pele muito escura aqui e na Europa pode sofrer algum viés, influenciando significativamente na classificação dos fenótipos. Por exemplo, dois indivíduos com tipos de pele consideradas no Brasil como “muito clara” e “intermediária” podem ser classificados como de peles “clara” e “escura” na Europa, respectivamente. Isso pode explicar parcialmente porque nenhum indivíduo foi predito com pele muito clara no presente estudo e porque todas as pessoas com pele muito escura foram preditas corretamente.

No artigo original do *Snipper10* para predição da cor da pele (Maroñas et al. 2014), o conjunto de 167 indivíduos utilizado para teste da ferramenta (utilizando classificador Bayesiano e estabelecendo um *cut-off* de 3:1) resultou nos seguintes parâmetros de sensibilidade: 0,969 para pele branca, 0,620 para pele intermediária e 0,300 para pele escura, com uma sensibilidade total de 0,814 com 6,6% de indivíduos não-classificados. Apesar da sensibilidade para pele escura também ser baixa, os resultados mostrados no estudo original são um pouco melhores do que os obtidos aqui. O estudo original do *HIrisPlex-S* (Walsh et al. 2017) também testou a ferramenta *Snipper10*, usando o mesmo critério de classificação de pele (valor de L^*_{ab}), com o classificador Bayesiano e sem utilizar *cut-off* 3:1. Os seguintes parâmetros de AUC e sensibilidade foram obtidos, respectivamente: 0,79 e 0,97 para pele branca, 0,63 e 0,37 para pele intermediária e 0,64 e 0,30 para pele escura. Em nosso estudo, valores de AUC superiores foram obtidos, porém só a cor intermediária atingiu uma sensibilidade superior (0,507) à do estudo de Walsh et al. (2017).

Enquanto para o *HIrisPlex-S* temos sensibilidade 1 para pele muito escura e 0 para pele muito clara, é notável que o inverso ocorre com o *Snipper*: a ferramenta não mostra um bom desempenho na determinação de pele escura, porém é muito eficiente na determinação de pele branca. A sensibilidade total, apesar de não ser totalmente satisfatória, é maior que a apresentada pelo *HIrisPlex-S* (0,811 contra 0,389). Isto pode ser explicado em parte por a divisão de cores de pele do *Snipper10* ser mais simples, apresentando apenas 3 categorias. Adicionalmente, um número maior de amostras não-europeias de diferentes populações

(incluindo amostras sul americanas) foi utilizado no conjunto de indivíduos empregados pelo *Snipper10* para desenvolver seu modelo estatístico (46,3% contra 16,2% do *HIrisPlex-S*), o que pode contribuir para que seus resultados de predição sejam mais acurados, independentemente da região geográfica e composição genética da população em estudo.

É difícil comparar os resultados de sensibilidade obtidos por *Snipper10* e *HIrisPlex-S* devido ao fato de os esquemas de classificação usados para definir os tons de pele serem diferentes. Podemos tentar aproximar os dois esquemas, porém a comparação entre eles ainda é difícil. Para a análise com o *Snipper*, é notável que a maioria (70,39%) dos indivíduos foi classificada com pele branca de acordo com o parâmetro L^*_{ab} . No estudo de Walsh, que também avalia a performance do *Snipper10*, é sugerido que as categorias “branca”, “intermediária” e “escura” do *Snipper* correspondem aproximadamente às categorias “muito clara/clara/intermediária”, “escura” e “muito escura”, do *HIrisPlex-S* (Walsh et al. 2017), respectivamente. Levando esse fato em consideração, em nosso estudo temos 425 indivíduos (85,17%) nas categorias “muito clara” a “intermediária”, ou seja, que deveriam ter sido classificados como brancos pelo parâmetro L^*_{ab} , 49 indivíduos na categoria “escura”, que deveriam ter sido classificados como intermediários e 25 na categoria “muito escura”, os quais seriam classificados como escuros. Isso nos mostra que, mesmo com a aproximação sugerida, ainda assim alguns indivíduos apresentam classificações diferentes do esperado. Enquanto os indivíduos são classificados pelo esquema Fitzpatrick a partir da cor do seu rosto e braços, áreas com maior exposição ao sol, além da capacidade de bronzeamento do indivíduo, o esquema de classificação usado pelo *Snipper10* utiliza o valor de L^*_{ab} correspondente à parte interior do braço, a qual tem baixa exposição solar. Já foram apresentados previamente informações do nosso banco de dados que mostram que tomando por base a parte interior do braço, os indivíduos são classificados como mais claros do que se tomássemos como referência a cor do rosto, pescoço e braços. Assim, tentando aproximar os dois esquemas de classificação, devido à diferença das áreas utilizadas para classificação, os indivíduos na verdade poderiam ser classificados em tons ainda mais claros utilizando o parâmetro L^*_{ab} , o que nos leva a supor que deveríamos ter um número maior do que 70,39% ou até mesmo 85,17% de indivíduos classificados como “brancos” e um número menor de indivíduos classificados como “escuros”. Como isso não ocorre, fica demonstrado a dificuldade em se entender a relação entre os dois métodos de classificação.

7.3.4 – Cor da pele, olhos e cabelos

Os desenvolvedores da ferramenta *HIrisPlex-S* alegam que ela pode ser usada em países localizados fora da Europa (Walsh et al. 2017; Chaitanya et al. 2018). Olhando pelo lado da aplicabilidade forense desse sistema no Brasil, os resultados combinados devem ser levados em consideração. Combinando as três predições (olho, cabelo e pele) obtidas para cada um dos indivíduos, tem-se que em apenas 26,9% dos casos (*threshold* de $p > 0,7$) as três estavam simultaneamente corretas, desclassificando 11,45% dos indivíduos. Esse dado produz sérios questionamentos quanto à aplicação forense do *HIrisPlex-S* no Brasil, pois em uma situação real espera-se que ao usar a ferramenta as predições combinadas apresentem acurácia elevada, gerando um perfil confiável do suspeito.

Combinando as predições do *Snipper23*, *Snipper12* e *Snipper10* para as três características, isto é, olhos, cabelos e pele, respectivamente, aplicando o *cut-off* e desconsiderando a fase haplotípica dos SNPs rs12913832 e rs1129038 em todos eles, tem-se que em apenas 29,8% dos casos todas as predições estavam corretas. Este valor, apesar de ser um pouco maior que aquele obtido pelo sistema *HIrisPlex-S* (26,9%), continua sendo insatisfatório e mostra a inviabilidade de uso dessa ferramenta na população brasileira. Além disso, combinando os resultados das três ferramentas, 46,51% dos indivíduos não são classificados, contra 11,45% pelo *HIrisPlex-S*.

Uma observação a ser feita é a de que quanto maior a estratificação de uma categoria fenotípica, mais complexa se torna a atividade de predição, dificultando assim que a mesma atinja performance acurada em todas as subcategorias. Por exemplo, o sistema *HIrisPlex-S* conta com cinco divisões para a cor da pele, enquanto o *Snipper12* conta com apenas três. Apesar dos esquemas de classificação serem diferentes, é perceptível que uma menor divisão de categorias facilita a classificação dos indivíduos pela ferramenta. Assim, para uma boa aplicação na rotina forense, o ideal seria que as características fenotípicas fossem estratificadas de forma simples, contando com 3 ou 4 categorias.

7.4 ASSOCIAÇÕES FENOTÍPICAS

Realizando uma regressão logística, foram encontradas 148 associações alélicas ao todo ($OR > 5$). Porém, apenas 38 dos 61 marcadores pertencentes às ferramentas de predição fenotípicas apresentaram alguma associação com os fenótipos de pigmentação (Apêndice B). Se considerarmos também as associações fracas ($1 < OR < 5$), temos 351 associações, mostradas no Apêndice C. Ainda assim, existem sete marcadores que não mostram associações com nenhuma característica de pigmentação. Esse dado nos mostra que muitos dos SNPs utilizados para predição de fenótipos na Europa não possuem a mesma eficiência quando utilizados para este fim no Brasil. Diante disso, pode-se supor que o uso deles nos sistemas de predição pode interferir negativamente na performance de tais ferramentas em populações miscigenadas, pois ao se basearem em associações entre genótipos e fenótipos que não são válidas nessas populações, podem levar a predições equivocadas.

Adicionando a ancestralidade individual como fator de confusão, foram encontradas apenas 26 associações entre 20 alelos (Tabela 17). Uma associação a se destacar é a do alelo T do marcador rs1805007 com cabelos ruivos, a qual apresentou *Odds Ratio* de 412,8. Esse alto valor pode ser explicado devido ao fato do alelo T ser muito raro em nossa população (frequência de 0,004). Apesar do número de associações encontradas ser baixo, é importante lembrar que a ancestralidade está relacionada com a pigmentação em muitos casos, mesmo que a ancestralidade de um indivíduo nem sempre se reflita em seus fenótipos, como já mencionado anteriormente na Introdução. Por exemplo, sabe-se que o SNP rs12913832 tem papel fundamental na determinação de olhos castanhos (alelos AA) e azuis (alelos GG) (Liu et al. 2009). Porém, esse SNP também está fortemente relacionado com ancestralidade, sendo um dos marcadores pertencentes ao sistema *34-Plex* para inferência de ancestralidade. O mesmo ocorre nos casos dos SNPs rs16891982 e rs1426654, também pertencentes ao sistema *34-Plex*: ambos apresentavam oito associações com $OR > 5$ e ao adicionarmos a ancestralidade não restou nenhuma associação. Outro estudo realizado em população brasileira também já apresentou associações encontradas entre esses SNPs e índice de melanina (Cerqueira et al. 2014). Se é sabido que esses marcadores estão relacionados com pigmentação, porém as associações são perdidas com a adição da ancestralidade no modelo, esse ajuste não demonstra a total realidade. Se utilizamos marcadores de pigmentação para estimar a ancestralidade, estaremos indiretamente usando a pigmentação para classificar a ancestralidade também, o que influencia na hora de adicionarmos essa informação como fator de confusão.

Talvez, para que essas associações importantes não fossem perdidas, o ideal fosse utilizar um conjunto de marcadores para inferência de ancestralidade que não possuam marcadores relacionados também com pigmentação.

Observando apenas as associações mostradas na Tabela 17, válidas para um indivíduo independentemente de sua origem biogeográfica, podemos fazer algumas observações: o SNP rs2238289 possui poder de diferenciação entre olhos castanhos e azuis e intermediários e azuis, porém não é usado para predição desse fenótipo por nenhum dos conjuntos. Inclusive, ele mostra um alto OR na diferenciação de olhos intermediários e azuis, que é o grande problema das ferramentas preditivas. Já em relação ao cabelo, dos 16 marcadores associados a esse fenótipo, apenas sete pertencem às ferramentas de predição dessa característica. Em relação à cor da pele, apenas um dos quatro marcadores que mostraram associação não é utilizado por conjuntos de predição da cor da pele. Mesmo assim, não existem bons marcadores para diferenciar a pele escura das outras cores de pele. Novamente, isso pode ser também devido ao fato do tom da pele também estar relacionado à ancestralidade, assim como a fatores ambientais, como a exposição à luz solar.

7.5 ANCESTRALIDADE GENÔMICA

Os conjuntos de marcadores, com exceção ao *Snipper10*, não se mostraram úteis na inferência da ancestralidade, não separando bem as populações parentais. Esse resultado é em parte esperado, seguindo a premissa de que a maioria dos marcadores está relacionado à pigmentação e é independente da origem biogeográfica de um indivíduo. Enquanto que a variação na pigmentação da pele apresenta boa correlação com a afiliação biogeográfica do indivíduo, o que explica a boa performance do sistema *Snipper10*, a variação na pigmentação dos olhos e cabelos é praticamente restrita a populações europeias. Adicionalmente, muitos dos marcadores empregados não apresentam elevado nível de polimorfismo e, apesar de um estudo realizado anteriormente em população brasileira afirmar que com 12 marcadores bem selecionados já é possível se estimar ancestralidade (Coelho et al. 2015), outro estudo sugere um painel de 192 SNPs para inferência da ancestralidade em população brasileira (Santos et al. 2016). De qualquer forma, tais marcadores foram muito bem estudados e escolhidos criteriosamente com intuito de estimar ancestralidade, diferentemente dos marcadores que

compõem as ferramentas preditivas aqui estudadas. Mesmo ao realizarmos a análise com o número total de marcadores, ainda assim podemos notar que as populações parentais europeia e asiática não são bem agrupadas em *clusters*, levando a conclusão que a má performance na inferência de ancestralidade é resultado dos marcadores utilizados, e não do número de marcadores empregados.

A ferramenta que apresenta o pior resultado é o *HLrisPlex-S*, pois o *software* não consegue separar a população asiática e africana em dois *clusters* distintos (Figura 7). Esse conjunto preditivo é o que possui o maior número de marcadores (41) dentre os seis, e a confusão gerada pode estar atribuída também ao fato de possuir marcadores ligados à três fenótipos diferentes: cor dos olhos, cabelos e pele. Além disso, muitos marcadores estão concentrados em um mesmo gene (11 deles no *MC1R* e 10 nos genes *OCA2/HERC2*, por exemplo), podendo, portanto, apresentar elevado nível de desequilíbrio de ligação. Já observando os gráficos triangulares obtidos para os *Snipper7*, *13* e *23*, podemos notar que não foi possível agrupar bem a população europeia em um dos vértices do triângulo (Figura 7). Além disso, a adição de mais marcadores proporciona uma melhor separação entre os *clusters* asiático e africano. Em relação ao *Snipper12*, podemos notar que o *software* não conseguiu separar muito bem as populações africana e asiática. Isso pode ser em parte explicado pois os marcadores desse conjunto têm como objetivo diferenciar cores de cabelos e ambas as populações apresentam em sua maioria cabelos pretos. Por fim, o *Snipper10* foi o único conjunto que conseguiu separar bem as populações parentais, apresentando estimativas de ancestralidade europeia, africana e asiática que refletem melhor a miscigenação população de Ribeirão Preto: 0,705, 0,158 e 0,137.

No estudo realizado por Salzano e Sans (2014) foram apresentados dados de ancestralidade para a região Sudeste a partir de conjuntos de 12 STRs, 28 e 40 AIMs, com proporções de ancestralidade europeia, africana e ameríndia variando de 0,74 a 0,80, 0,14 a 0,19 e 0,06 a 0,07, respectivamente. No presente estudo obteve-se uma menor contribuição europeia e ameríndia, enquanto a contribuição africana permaneceu dentro da variação apresentada.

Já no estudo realizado por Valle-Silva et. al (2019) utilizando as mesmas amostras aqui estudadas da população de Ribeirão Preto e as populações parentais do projeto *1000 Genomes*, o *software* Structure definiu as seguintes proporções: 0,718 (EUR), 0,215 (AFR) e 0,060 (EAS). Esse estudo é o que apresenta metodologia mais próxima à utilizada aqui. Comparando os valores obtidos, pode-se perceber que a contribuição europeia obtida no presente estudo é bem

próxima ao valor obtido no estudo de Valle-Silva. Porém, a estimativa de contribuição asiática obtida em nosso estudo parece estar inflacionada em relação aos valores apontados por Valle-Silva, causando assim uma diminuição na contribuição africana. Ao discutirmos esse fato, é válido lembrar que, embora o conjunto *Snipper10* tenha conseguido separar de forma adequada as populações parentais, os marcadores não foram selecionados com este objetivo.

Conclusão

8. Conclusão

A imputação foi realizada com sucesso, e os dois programas utilizados se mostraram muito eficientes, possuindo cada um suas vantagens e desvantagens, sendo difícil dizer qual deles é o melhor. Portanto, a unificação dos resultados gerados pelos dois métodos se mostrou uma opção excelente, maximizando a acurácia obtida com um número baixo de genótipos indefinidos.

De maneira geral, as ferramentas não apresentaram performance boa o suficiente para serem usadas com confiabilidade na população brasileira. A maior falha das ferramentas preditivas reside na classificação dos fenótipos intermediários (olhos verde/mel, cabelos loiro-escuro/castanho-claro e pele intermediária). Provavelmente isso ocorre pelo fato do Brasil possuir uma população muito miscigenada, ao contrário da população europeia, onde as ferramentas foram desenvolvidas.

As análises de associação revelam que muitos marcadores não possuem relação significativa com fenótipos de pigmentação. Por outro lado, existem marcadores associados com características diferentes daquelas para as quais a utilização foi originalmente proposta. Dessa maneira, este trabalho ressalta que, apesar de já existirem ferramentas de predição, elas devem ser atualizadas para o uso em populações miscigenadas, por meio (a) da inclusão dos marcadores que já se mostraram associados com outras características até o momento e retirar aqueles que não estão associados com pigmentação, e (b) da busca por marcadores associados com fenótipos de pigmentação em populações miscigenadas. Isso é de extrema importância para que se possa criar uma ferramenta de maior valor preditivo e que possa ser usada com confiabilidade em tais populações.

Também foi possível concluir que as ferramentas não são úteis na inferência da ancestralidade, com exceção do conjunto *Snipper10*. Isso é esperado pois, teoricamente, os marcadores presentes nas ferramentas independem da origem biogeográfica do indivíduo, sendo que apenas a variação na pigmentação da pele apresenta boa correlação com a afiliação biogeográfica da amostra analisada.

Referências

- Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque J-C, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuña-Alonzo V, et al. 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat Commun.* 7(1):10815. doi:10.1038/ncomms10815.
- Ambrosio AL, Boyle JA, Aradi AE, Christian KA, Di Pietro SM. 2016. TPC2 controls pigmentation by regulating melanosome pH and size. *Proc Natl Acad Sci.* 113(20):5622–5627. doi:10.1073/pnas.1600108113.
- Andrade ES, Fracasso NCA, Strazza Júnior PS, Simões AL, Mendes-Junior CT. 2017. Associations of OCA2 - HERC2 SNPs and haplotypes with human pigmentation characteristics in the Brazilian population. *Leg Med.* 24:78–83. doi:10.1016/j.legalmed.2016.12.003.
- Bellono NW, Escobar IE, Oancea E. 2016. A melanosomal two-pore sodium channel regulates pigmentation. *Sci Rep.* 6(1):26570. doi:10.1038/srep26570.
- Biasutti R. 1953. *Le razze e i popoli della terra*. Torino: Unione Tipografico Editrice Torinese.
- Branicki W, Liu F, van Duijn K, Draus-Barini J, Pośpiech E, Walsh S, Kupiec T, Wojas-Pelc A, Kayser M. 2011. Model-based prediction of human hair color using DNA variants. *Hum Genet.* 129:443–454. doi:10.1007/s00439-010-0939-8.
- Castelli EC, Mendes-Junior CT, Sabbagh A, Porto IOP, Garcia A, Ramalho J, Lima THA, Massaro JD, Dias FC, Collares CVA, et al. 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. doi:10.1016/j.humimm.2015.06.016.
- Cerqueira CCS de, Hünemeier T, Gomez-Valdés J, Ramallo V, Volasko-Krause CD, Barbosa AAL, Vargas-Pinilla P, Dornelles RC, Longo D, Rothhammer F, et al. 2014. Implications of the Admixture Process in Skin Color Molecular Assessment. Palau F, editor. *PLoS One.* 9(5):e96886. doi:10.1371/journal.pone.0096886.
- Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pośpiech E, Kukla-Bartoszek M, Sijen T, Knijff P de, Liu F, Branicki W, et al. 2018. The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Sci Int Genet.* 35:123–135. doi:10.1016/j.fsigen.2018.04.004.

- Cichorek M, Wachulska M, Stasiewicz A, Tymińska A. 2013. Skin melanocytes: biology and development. *Adv Dermatology Allergol.* 1(1):30–41. doi:10.5114/pdia.2013.33376.
- Coelho AVC, Moura RR, Cavalcanti CAJ, Guimarães RL, Sandrin-Garcia P, Crovella S, Brandão LAC. 2015. A rapid screening of ancestry for genetic association studies in an admixed population from Pernambuco, Brazil. *Genet Mol Res.* 14(1):2876–2884. doi:10.4238/2015.March.31.18.
- Cook AL, Chen W, Thurber AE, Smit DJ, Smith AG, Bladen TG, Brown DL, Duffy DL, Pastorino L, Bianchi-Scarra G, et al. 2009. Analysis of Cultured Human Melanocytes Based on Polymorphisms within the SLC45A2/MATP, SLC24A5/NCKX5, and OCA2/P Loci. *J Invest Dermatol.* 129(2):392–405. doi:10.1038/jid.2008.211.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics.* 27(15):2156–2158. doi:10.1093/bioinformatics/btr330.
- Dario P, Mouriño H, Oliveira AR, Lucas I, Ribeiro T, Porto MJ, Costa Santos J, Dias D, Corte Real F. 2015. Assessment of IrisPlex-based multiplex for eye and skin color prediction with application to a Portuguese population. *Int J Legal Med.* 129(6):1191–1200. doi:10.1007/s00414-015-1248-5.
- Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. 2016. Next-generation genotype imputation service and methods. *Nat Genet.* 48(10):1284–1287. doi:10.1038/ng.3656.
- de Araújo Lima F, de Toledo Gonçalves F, Fridman C. 2015. SLC24A5 and ASIP as phenotypic predictors in Brazilian population for forensic purposes. *Leg Med.* 17(4):261–266. doi:10.1016/j.legalmed.2015.03.001.
- Delaneau O, Marchini J, Zagury J. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 9(2):179–181. doi:10.1038/nmeth.1785.
- Delaneau O, Zagury J, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 10(1):5–6. doi:10.1038/nmeth.2307.
- Dembinski GM, Picard CJ. 2014. Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population. *Forensic Sci Int Genet.* 9(1):111–117. doi:10.1016/j.fsigen.2013.12.003.

- DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43(5):491–498. doi:10.1038/ng.806.
- Ferreira LB, Mendes-Junior CT, Wiezel CEV, Luizon MR, Simões AL. 2006. Genomic ancestry of a sample population from the state of São Paulo, Brazil. *Am J Hum Biol.* 18(5):702–705. doi:10.1002/ajhb.20474.
- Fitzpatrick TB. 1988. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Arch Dermatol.* 124(6):869–871. doi:10.1001/archderm.1988.01670060015008.
- Fondevila M, Phillips C, Santos C, Freire Aradas A, Vallone PM, Butler JM, Lareu MV, Carracedo Á. 2013. Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int Genet.* 7(1):63–74. doi:10.1016/j.fsigen.2012.06.007.
- Fracasso NC de A, de Andrade ES, Wiezel CEV, Andrade CCF, Zañão LR, da Silva MS, Marano LA, Donadi EA, C. Castelli E, Simões AL, et al. 2017. Haplotypes from the SLC45A2 gene are associated with the presence of freckles and eye, hair and skin pigmentation in Brazil. *Leg Med.* 25:43–51. doi:10.1016/j.legalmed.2016.12.013.
- Freire-Aradas A, Ruiz Y, Phillips C, Maroñas O, Söchtig J, Tato AG, Dios JÁ, de Cal MC, Silbiger VN, Luchessi AD, et al. 2014. Exploring iris colour prediction and ancestry inference in admixed populations of South America. *Forensic Sci Int Genet.* 13:3–9. doi:10.1016/j.fsigen.2014.06.007.
- Goodwin W, Linacre A, Hadi S. 2011. *An Introduction to Forensic Genetics*. 2nd ed. John Wiley & Sons Ltd.
- Graf J, Hodgson R, van Daal A. 2005. Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation. *Hum Mutat.* 25(3):278–284. doi:10.1002/humu.20143.
- Graffelman J, Nelson S, Gogarten SM, Weir BS. 2015. Exact Inference for Hardy-Weinberg Proportions with Missing Genotypes: Single and Multiple Imputation. *G3 Genes, Genomes, Genet.* 5(11):2365–2373. doi:10.1534/g3.115.022111.
- Hill HZ. 1992. The function of melanin or six blind people examine an elephant. *BioEssays.*

- 14(1):49–56. doi:10.1002/bies.950140111.
- Howie B, Marchini J, Stephens M. 2011. Genotype Imputation with Thousands of Genomes. *G3 Genes, Genomes, Genet.* 1(6):457–470. doi:10.1534/g3.111.001198.
- Howie BN, Donnelly P, Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Schork NJ, editor. *PLoS Genet.* 5(6):e1000529. doi:10.1371/journal.pgen.1000529.
- Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Mol Ecol Resour.* 9(5):1322–1332. doi:10.1111/j.1755-0998.2009.02591.x.
- Kastelic V, Pošpiech E, Draus-Barini J, Branicki W, Drobnič K. 2013. Prediction of eye color in the Slovenian population using the IrisPlex SNPs. *Croat Med J.* 54(4):381–386. doi:10.3325/cmj.2013.54.381.
- Kayser M. 2015. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci Int Genet.* 18:33–48. doi:10.1016/j.fsigen.2015.02.003.
- Kayser M, de Knijff P. 2011. Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet.* 12(3):179–192. doi:10.1038/nrg2952.
- Kayser M, 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. doi:10.1016/j.fsigen.2009.01.012.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 25(14):1754–1760. doi:10.1093/bioinformatics/btp324.
- Lima THA, Buttura RV, Donadi EA, Veiga-Castelli LC, Mendes-Junior CT, 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. doi:10.1016/j.humimm.2016.07.231.
- Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens ACJW, Kayser M. 2009. Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol.* 19(5):R192–R193. doi:10.1016/j.cub.2009.01.027.

- Liu F, Wen B, Kayser M. 2013. Colorful DNA polymorphisms in humans. *Semin Cell Dev Biol.* 24(6–7):562–575. doi:10.1016/j.semcdb.2013.03.013.
- Loomis WF. 1967. Skin-Pigment Regulation of Vitamin-D Biosynthesis in Man: Variation in solar ultraviolet at different latitudes may have caused racial differentiation in man. *Science* (80-). 157(3788):501–506. doi:10.1126/science.157.3788.501.
- Marcorin L. 2017. Avaliação da variabilidade genética do gene *MITF* e suas associações com fenótipos de pigmentação em amostra da população brasileira. [Ribeirão Preto]: Universidade de São Paulo.
- Maroñas O, Phillips C, Söchtig J, Gomez-Tato A, Cruz R, Alvarez-Dios J, de Cal MC, Ruiz Y, Fondevila M, Carracedo Á, et al. 2014. Development of a forensic skin colour predictive test. *Forensic Sci Int Genet.* 13:34–44. doi:10.1016/j.fsigen.2014.06.017.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* 17(1):10. doi:10.14806/ej.17.1.200.
- 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. doi:10.1101/gr.107524.110.
- Miller SA, Dykes DD, Polesky HF. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* 16(3):1215.
- Muniz YCN, Ferreira LB, Mendes-Junior CT, Wiezel CEV, Simões AL. 2008. Genomic ancestry in urban Afro-Brazilians. *Ann Hum Biol.* 35(1):104–111. doi:10.1080/03014460701782256.
- Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD. 2006. Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Mol Biol Evol.* 24(3):710–722. doi:10.1093/molbev/msl203.
- Oliveira MLG de. 2017. Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira. [Ribeirão Preto]: Universidade de São Paulo.
- Oliveira MLG de, Veiga-Castelli LC, Marcorin L, Debortoli G, Pereira ALE, Fracasso NC de

- A, Silva G do V, Souza AS, Massaro JD, Simões AL, et al. 2018. Extended HLA-G genetic diversity and ancestry composition in a Brazilian admixed population sample: Implications for HLA-G transcriptional control and for case-control association studies. *Hum Immunol.* 79(11):790–799. doi:10.1016/j.humimm.2018.08.005.
- Parra EJ. 2007. Human pigmentation variation: Evolution, genetic basis, and implications for public health. *Am J Phys Anthropol.* 134(S45):85–105. doi:10.1002/ajpa.20727.
- Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM, Pena SDJ. 2003. Color and genomic ancestry in Brazilians. *Proc Natl Acad Sci.* 100(1):177–182. doi:10.1073/pnas.0126614100.
- Pena SDJ, Bastos-Rodrigues L, Pimenta JR, Bydlowski SP. 2009. DNA tests probe the genomic ancestry of Brazilians. *Brazilian J Med Biol Res.* 42(10):870–876. doi:10.1590/S0100-879X2009005000026.
- Pena SDJ, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy F de SG, Kohlrausch F, Magno LAV, Montenegro RC, Moraes MO, et al. 2011. The Genomic Ancestry of Individuals from Different Geographical Regions of Brazil Is More Uniform Than Expected. *PLoS One.* 6(2):e17063. doi:10.1371/journal.pone.0017063.
- Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Álvarez-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. doi:10.1016/j.fsigen.2007.06.008.
- Pimenta JR, Zuccherato LW, Debes AA, Maselli L, Soares RP, Moura-Neto RS, Rocha J, Bydlowski SP, Pena SDJ. 2006. Color and Genomic Ancestry in Brazilians: A Study with Forensic Microsatellites. *Hum Hered.* 62(4):190–195. doi:10.1159/000096872.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics.* 155(2):945–959.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.* 81(3):559–575. doi:10.1086/519795.
- Relethford JH. 1997. Hemispheric difference in human skin color. *Am J Phys Anthropol.*

104(4):449–457. doi:10.1002/(SICI)1096-8644(199712)104:4<449::AID-AJPA2>3.0.CO;2-N.

- Robins AH. 1991. Biological perspectives on human pigmentation. Cambridge, NY: Cambridge University Press.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol.* 29(1):24–26. doi:10.1038/nbt.1754.
- Rodrigues de Moura R, Coelho AVC, de Queiroz Balbino V, Crovella S, Brandão LAC. 2015. Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. *Am J Hum Biol.* 27(5):674–680. doi:10.1002/ajhb.22714.
- Rousset F. 2008. Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour.* 8(1):103–106. doi:10.1111/j.1471-8286.2007.01931.x.
- Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, Casares de Cal M, Cruz R, Maroñas O, Söchtig J, Fondevila M, Rodriguez-Cid MJ, et al. 2013. Further development of forensic eye color predictive tests. *Forensic Sci Int Genet.* 7(1):28–40. doi:10.1016/j.fsigen.2012.05.009.
- Salvoro C, Faccinetto C, Zucchelli L, Porto M, Marino A, Occhi G, de los Campos G, Vazza G. 2019. Performance of four models for eye color prediction in an Italian population sample. *Forensic Sci Int Genet.* 40(March):192–200. doi:10.1016/j.fsigen.2019.03.008.
- Salzano FM, Sans M. 2014. Interethnic admixture and the evolution of Latin American populations. *Genet Mol Biol.* 37(1 suppl 1):151–170. doi:10.1590/S1415-47572014000200003.
- Santos HC, Horimoto AVR, Tarazona-Santos E, Rodrigues-Soares F, Barreto ML, Horta BL, Lima-Costa MF, Gouveia MH, Machado M, Silva TM, et al. 2016. A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: the Brazilian set. *Eur J Hum Genet.* 24(5):725–731. doi:10.1038/ejhg.2015.187.
- Santos NPC, Ribeiro-Rodrigues EM, Ribeiro-dos-Santos ÂKC, Pereira R, Gusmão L, Amorim A, Guerreiro JF, Zago MA, Matte C, Hutz MH, et al. 2010. Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-

- informative marker (AIM) panel. *Hum Mutat.* 31(2):184–190. doi:10.1002/humu.21159.
- Söchtig J, Phillips C, Maroñas O, Gómez-Tato A, Cruz R, Alvarez-Dios J, de Cal M-ÁC, Ruiz Y, Reich K, Fondevila M, et al. 2015. Exploration of SNP variants affecting hair colour prediction in Europeans. *Int J Legal Med.* 129(5):963–975. doi:10.1007/s00414-015-1226-y.
- Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, Kovacevic L, Marjanovic D, Caragine T, Prinz M, Wurmbach E. 2011. Prediction of eye and skin color in diverse populations using seven SNPs. *Forensic Sci Int Genet.* 5(5):472–478. doi:10.1016/j.fsigen.2010.10.005.
- Stephens M, Donnelly P. 2003. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *Am J Hum Genet.* 73(5):1162–1169. doi:10.1086/379378.
- Stephens M, Smith NJ, Donnelly P. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. *Am J Hum Genet.* 68(4):978–989. doi:10.1086/319501.
- Sturm RA. 2009. Molecular genetics of human pigmentation diversity. *Hum Mol Genet.* 18(R1):R9–R17. doi:10.1093/hmg/ddp003.
- Sturm RA, Box NF, Ramsay M. 1998. Human pigmentation genetics: the difference is only skin deep. *BioEssays.* 20(9):712–721. doi:10.1002/(SICI)1521-1878(199809)20:9<712::AID-BIES4>3.0.CO;2-I.
- Sturm RA, Teasdale RD, Box NF. 2001. Human pigmentation genes: identification, structure and consequences of polymorphic variation. *Gene.* 277(1–2):49–62. doi:10.1016/S0378-1119(01)00694-1.
- Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G, et al. 2008. Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet.* 40(7):835–837. doi:10.1038/ng.160.
- Telles EE. 1992. Residential Segregation by Skin Color in Brazil. *Am Sociol Rev.* 57(2):186. doi:10.2307/2096204.
- The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature.* 526(7571):68–74. doi:10.1038/nature15393.

- Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 14(2):178–192. doi:10.1093/bib/bbs017.
- Valle-Silva G do, Souza FDN de, Marcorin L, Pereira ALE, Carratto TMT, Debortoli G, Oliveira MLG de, Fracasso NC de A, Andrade ES de, Donadi EA, et al. 2019. Applicability of the SNPforID 52-plex panel for human identification and ancestry evaluation in a Brazilian population sample by next-generation sequencing. *Forensic Sci Int Genet.* 40:201–209. doi:10.1016/j.fsigen.2019.03.003.
- Van der Auwera GA, Carneiro Mauricio O., Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: *Current Protocols in Bioinformatics*. Vol. 11. Hoboken, NJ, USA: John Wiley & Sons, Inc. p. 11.10.1-11.10.33.
- Walsh S, Chaitanya L, Breslin K, Muralidharan C, Bronikowska A, Pospiech E, Koller J, Kovatsi L, Wollstein A, Branicki W, et al. 2017. Global skin colour prediction from DNA. *Hum Genet.* 136(7):847–863. doi:10.1007/s00439-017-1808-5.
- Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, Maeda H, Ishikawa T, Sijen T, de Knijff P, et al. 2014. Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. *Forensic Sci Int Genet.* 9:150–161. doi:10.1016/j.fsigen.2013.12.006.
- Walsh S, Lindenbergh A, Zuniga SB, Sijen T, de Knijff P, Kayser M, Ballantyne KN. 2011. Developmental validation of the IrisPlex system: Determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet.* 5(5):464–471. doi:10.1016/j.fsigen.2010.09.008.
- Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M. 2011. IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet.* 5(3):170–180. doi:10.1016/j.fsigen.2010.02.004.
- Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz A, Branicki W, Kayser M. 2013. The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet.* 7(1):98–115. doi:10.1016/j.fsigen.2012.07.005.
- Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L,

Topouzis F, Vingerling JR, et al. 2012. DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci Int Genet.* 6(3):330–340. doi:10.1016/j.fsigen.2011.07.009.

Weatherall IL, Coombs BD. 1992. Skin Color Measurements in Terms of CIELAB Color Space Values. *J Invest Dermatol.* 99(4):468–473. doi:10.1111/1523-1747.ep12616156.

Yun L, Gu Y, Rajeevan H, Kidd KK. 2014. Application of six IrisPlex SNPs and comparison of two eye color prediction systems in diverse Eurasia populations. *Int J Legal Med.* 128(3):447–453. doi:10.1007/s00414-013-0953-1.

Apêndice

APÊNDICE A

Tabela A1 – Populações parentais utilizadas para cálculos de ancestralidade, obtidas a partir dos dados do projeto *1000 Genomes*.

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	98	Africana
Serra Leoa – Mende	MSL	85	Africana
China – Han de Pequim	CHB	103	Asiática
Japão – Tóquio	JPT	104	Asiática
China – Dai de Xishuangbanna	CDX	93	Asiática
China – Han do Sul da China	CHS	105	Asiática
Vietnã – Kinh in Ho Chi Minh	KHV	99	Asiática
Total de indivíduos		1411	

APÊNDICE B

Tabela B1 – Associações alélicas encontradas para os 61 marcadores dos seis conjuntos preditivos analisados.

Marcador	Sistema Preditivo*	Alelo	Característica	<i>p</i> exato de Fisher	<i>Odds Ratio</i>	IC 95%
rs1129038	C, D, E, F, G	T	Olho Azul vs. Castanho	$3,86 \times 10^{-15}$	140,30	40,90-481,60
rs12913832	A, B, C, D, E, F, G	G	Olho Azul vs. Castanho	$3,11 \times 10^{-15}$	160,00	45,34-565,00
rs11636232	E, F	T	Olho Azul vs. Castanho	$1,24 \times 10^{-9}$	6,68	3,62-12,33
rs1129038	C, D, E, F, G	T	Olho Azul vs. não Azul	$4,99 \times 10^{-12}$	13,97	6,61-29,53
rs12913832	A, B, C, D, E, F, G	G	Olho Azul vs. não Azul	$7,96 \times 10^{-12}$	15,82	7,17-34,90
rs2238289	C	G	Olho Intermediário vs. Azul	$2,41 \times 10^{-2}$	5,33	1,24-22,83
rs1129038	C, D, E, F, G	T	Olho Intermediário vs. Castanho	$3,02 \times 10^{-20}$	21,60	11,24-41,51
rs12913832	A, B, C, D, E, F, G	G	Olho Intermediário vs. Castanho	$2,43 \times 10^{-20}$	21,60	11,26-41,45
rs11636232	E, F	T	Olho Intermediário vs. Castanho	$2,83 \times 10^{-14}$	5,68	3,63-8,89
rs1129038	C, D, E, F, G	T	Olho Intermediário vs. não Intermediário	$1,80 \times 10^{-20}$	6,11	4,17-8,96
rs12913832	A, B, C, D, E, F, G	G	Olho Intermediário vs. não Intermediário	$1,69 \times 10^{-20}$	6,00	4,11-8,76
rs10777129	H	A	Olho Castanho vs. Azul	$9,08 \times 10^{-4}$	5,65	2,03-15,70
rs1448484	H	G	Olho Castanho vs. Azul	$4,17 \times 10^{-3}$	17,65	2,48-125,90
rs4778138	F, G	G	Olho Castanho vs. Azul	$6,30 \times 10^{-8}$	9,30	4,14-20,85
rs4778241	E, F	C	Olho Castanho vs. Azul	$5,21 \times 10^{-5}$	7,70	2,86-20,68
rs7495174	F, G	G	Olho Castanho vs. Azul	$3,70 \times 10^{-4}$	8,53	2,62-27,73
rs7183877	E, F	A	Olho Castanho vs. Azul	$2,34 \times 10^{-2}$	9,82	1,36-70,70
rs2238289	C	G	Olho Castanho vs. Azul	$1,07 \times 10^{-6}$	34,78	8,36-144,70
rs6497292	C	G	Olho Castanho vs. Azul	$4,58 \times 10^{-4}$	34,43	4,76-249,20
rs916977	F	T	Olho Castanho vs. Azul	$1,37 \times 10^{-8}$	30,69	9,41-100,10
rs1667394	C, E, F	C	Olho Castanho vs. Azul	$3,91 \times 10^{-9}$	35,60	10,84-116,90
rs12592730	F	A	Olho Castanho vs. Azul	$1,20 \times 10^{-2}$	12,78	1,75-93,30
rs1426654	C, H	G	Olho Castanho vs. Azul	$8,79 \times 10^{-5}$	7,12	2,67-18,98
rs1448484	H	G	Olho Castanho vs. Intermediário	$3,34 \times 10^{-6}$	6,99	3,08-15,87
rs7495174	F, G	G	Olho Castanho vs. Intermediário	$3,60 \times 10^{-7}$	5,40	2,82-10,35
rs2238289	C	G	Olho Castanho vs. Intermediário	$4,47 \times 10^{-13}$	5,68	3,55-9,10
rs6497292	C	G	Olho Castanho vs. Intermediário	$3,18 \times 10^{-9}$	5,92	3,28-10,66
rs916977	F	T	Olho Castanho vs. Intermediário	$8,54 \times 10^{-17}$	6,99	4,42-11,06
rs1667394	C, E, F	C	Olho Castanho vs. Intermediário	$3,86 \times 10^{-17}$	7,03	4,47-11,08
rs4778138	F, G	G	Olho Castanho vs. não Castanho	$1,73 \times 10^{-16}$	5,34	3,59-7,96
rs4778241	E, F	C	Olho Castanho vs. não Castanho	$3,70 \times 10^{-8}$	5,13	2,86-9,17
rs7495174	F, G	G	Olho Castanho vs. não Castanho	$1,28 \times 10^{-9}$	6,08	3,40-10,89

Marcador	Sistema Preditivo*	Alelo	Característica	<i>p</i> exato de Fisher	<i>Odds Ratio</i>	IC 95%
rs2238289	C	G	Olho Castanho vs. não Castanho	$1,43 \times 10^{-18}$	7,72	4,89-12,17
rs6497292	C	G	Olho Castanho vs. não Castanho	$1,24 \times 10^{-12}$	7,87	4,45-13,91
rs916977	F	T	Olho Castanho vs. não Castanho	$6,65 \times 10^{-23}$	9,06	5,84-14,04
rs1667394	C, E, F	C	Olho Castanho vs. não Castanho	$1,18 \times 10^{-23}$	9,29	6,01-14,36
rs8051733	C	G	Cabelo Ruivo vs. Loiro-Claro	$3,05 \times 10^{-3}$	11,43	2,28-57,31
rs7495174	F, G	G	Cabelo Ruivo vs. Castanho-Claro	$3,41 \times 10^{-3}$	5,02	1,70-14,77
rs12592730	F	A	Cabelo Ruivo vs. Castanho-Claro	$1,55 \times 10^{-3}$	7,17	2,12-24,3
rs12931267	G	G	Cabelo Ruivo vs. Castanho-Claro	$2,35 \times 10^{-4}$	97,15	8,48-1113,00
rs1805007	B, C, G	T	Cabelo Ruivo vs. Castanho-Claro	$4,11 \times 10^{-6}$	382,40	30,45-4803,00
rs1805008	B, C, G	T	Cabelo Ruivo vs. Castanho-Claro	$1,13 \times 10^{-3}$	11,62	2,65-50,88
rs8051733	C	G	Cabelo Ruivo vs. Castanho-Claro	$5,83 \times 10^{-5}$	9,46	3,16-28,29
rs12931267	G	G	Cabelo Ruivo vs. Castanho-Escuro	$4,59 \times 10^{-5}$	62,04	8,52-451,60
rs11547464	B, C, G	A	Cabelo Ruivo vs. Castanho-Escuro	$6,48 \times 10^{-3}$	17,78	2,24-141,10
rs1805007	B, C, G	T	Cabelo Ruivo vs. Castanho-Escuro	$3,24 \times 10^{-7}$	161,20	22,94-1133,00
rs1805008	B, C, G	T	Cabelo Ruivo vs. Castanho-Escuro	$3,37 \times 10^{-3}$	7,79	1,98-30,74
rs8051733	C	G	Cabelo Ruivo vs. Castanho-Escuro	$4,31 \times 10^{-5}$	9,21	3,18-26,67
rs1129038	C, D, E, F, G	T	Cabelo Ruivo vs. Preto	$7,55 \times 10^{-4}$	5,24	2,00-13,75
rs12913832	A, B, C, D, E, F, G	G	Cabelo Ruivo vs. Preto	$6,59 \times 10^{-4}$	5,31	2,03-13,88
rs12931267	G	G	Cabelo Ruivo vs. Preto	$3,21 \times 10^{-4}$	87,39	7,65-998,00
rs1805007	B, C, G	T	Cabelo Ruivo vs. Preto	$5,92 \times 10^{-6}$	346,40	27,58-4351,00
rs1805008	B, C, G	T	Cabelo Ruivo vs. Preto	$4,07 \times 10^{-4}$	65,71	6,46-668,90
rs8051733	C	G	Cabelo Ruivo vs. Preto	$8,55 \times 10^{-4}$	5,90	2,08-16,76
rs12931267	G	G	Cabelo Ruivo vs. não Ruivo	$6,07 \times 10^{-7}$	81,51	14,47-459,30
rs312262906	B	A	Cabelo Ruivo vs. não Ruivo	$1,02 \times 10^{-2}$	41,40	2,41-709,90
rs11547464	B, C, G	A	Cabelo Ruivo vs. não Ruivo	$7,20 \times 10^{-4}$	23,17	3,75-143,20
rs1805007	B, C, G	T	Cabelo Ruivo vs. não Ruivo	$2,29 \times 10^{-9}$	252,60	41,16-1551,00
rs1805008	B, C, G	T	Cabelo Ruivo vs. não Ruivo	$1,49 \times 10^{-4}$	12,76	3,42-47,59
rs8051733	C	G	Cabelo Ruivo vs. não Ruivo	$2,91 \times 10^{-5}$	7,97	3,01-21,08
rs1129038	C, D, E, F, G	T	Cabelo Loiro-Claro vs. Ruivo	$7,16 \times 10^{-3}$	26,56	2,43-289,80
rs12913832	A, B, C, D, E, F, G	G	Cabelo Loiro-Claro vs. Ruivo	$7,16 \times 10^{-3}$	26,56	2,43-289,80
rs11636232	E, F	T	Cabelo Loiro-Claro vs. Ruivo	$1,73 \times 10^{-2}$	9,29	1,48-58,21
rs1129038	C, D, E, F, G	T	Cabelo Loiro-Claro vs. Castanho-Claro	$5,32 \times 10^{-3}$	17,69	2,35-133,50
rs12913832	A, B, C, D, E, F, G	G	Cabelo Loiro-Claro vs. Castanho-Claro	$5,85 \times 10^{-3}$	17,09	2,27-128,60
rs1129038	C, D, E, F, G	T	Cabelo Loiro-Claro vs. Loiro-Escuro	$7,09 \times 10^{-5}$	64,36	8,25-502,10
rs12913832	A, B, C, D, E, F, G	G	Cabelo Loiro-Claro vs. Loiro-Escuro	$7,72 \times 10^{-5}$	62,62	8,05-487,00
rs11547464	B, C, G	A	Cabelo Loiro-Claro vs. Loiro-Escuro	$1,59 \times 10^{-2}$	12,31	1,60-94,64
rs35264875	G	T	Cabelo Loiro-Claro vs. Preto	$1,44 \times 10^{-3}$	5,68	1,95-16,53
rs12896399	A, C, D, E, F	T	Cabelo Loiro-Claro vs. Preto	$2,77 \times 10^{-4}$	5,82	2,25-15,03

Marcador	Sistema Preditivo*	Alelo	Característica	<i>p</i> exato de Fisher	Odds Ratio	IC 95%
rs1129038	C, D, E, F, G	T	Cabelo Loiro-Claro vs. Preto	$1,89 \times 10^{-5}$	147,10	14,94-1448,00
rs12913832	A, B, C, D, E, F, G	G	Cabelo Loiro-Claro vs. Preto	$2,08 \times 10^{-5}$	142,40	14,52-1396,00
rs11636232	E, F	T	Cabelo Loiro-Claro vs. Preto	$5,72 \times 10^{-5}$	8,83	3,06-25,50
rs1805009	B, G	C	Cabelo Loiro-Claro vs. Preto	$4,36 \times 10^{-2}$	8,14	1,06-62,44
rs1129038	C, D, E, F, G	T	Cabelo Loiro-Claro vs. não Loiro-Claro	$4,41 \times 10^{-4}$	34,55	4,79-249,10
rs12913832	A, B, C, D, E, F, G	G	Cabelo Loiro-Claro vs. não Loiro-Claro	$4,85 \times 10^{-4}$	33,39	4,65-239,20
rs11547464	B, C, G	A	Cabelo Loiro-Claro vs. não Loiro-Claro	$2,39 \times 10^{-3}$	15,88	2,67-94,66
rs1129038	C, D, E, F, G	T	Cabelo Castanho-Claro vs. Preto	$3,69 \times 10^{-16}$	15,80	5,42-15,80
rs12913832	A, B, C, D, E, F, G	G	Cabelo Castanho-Claro vs. Preto	$2,41 \times 10^{-16}$	16,19	5,53-16,19
rs2402130	B, C, H	G	Cabelo Castanho-Escuro vs. Ruivo	$2,80 \times 10^{-2}$	9,92	1,28-76,75
rs16891982	A, B, C, D, E, F, H	C	Cabelo Castanho-Escuro vs. Loiro-Claro	$1,31 \times 10^{-3}$	11,07	2,56-47,97
rs4778138	F, G	G	Cabelo Castanho-Escuro vs. Loiro-Claro	$3,74 \times 10^{-3}$	6,16	1,80-21,06
rs4778241	E, F	C	Cabelo Castanho-Escuro vs. Loiro-Claro	$2,62 \times 10^{-2}$	5,89	1,23-28,16
rs6119471	C, H	G	Cabelo Castanho-Escuro vs. Castanho-Claro	$1,20 \times 10^{-4}$	7,59	2,70-21,32
rs1805008	B, C, G	T	Cabelo Castanho-Escuro vs. Preto	$4,26 \times 10^{-2}$	8,43	1,07-66,26
rs16891982	A, B, C, D, E, F, H	C	Cabelo Preto vs. Ruivo	$6,36 \times 10^{-5}$	16,13	4,13-63,02
rs28777	B, C, G	C	Cabelo Preto vs. Ruivo	$4,59 \times 10^{-4}$	15,54	3,35-72,11
rs2402130	B, C, H	G	Cabelo Preto vs. Ruivo	$2,40 \times 10^{-2}$	10,38	1,36-79,21
rs4778232	E, F	T	Cabelo Preto vs. Ruivo	$1,77 \times 10^{-3}$	5,28	1,86-14,99
rs1448484	H	G	Cabelo Preto vs. Ruivo	$1,80 \times 10^{-2}$	11,33	1,52-84,72
rs8024968	E, F	T	Cabelo Preto vs. Ruivo	$1,83 \times 10^{-2}$	5,89	1,35-25,68
rs1470608	C	T	Cabelo Preto vs. Ruivo	$1,72 \times 10^{-3}$	5,94	1,95-18,09
rs1375164	F	T	Cabelo Preto vs. Ruivo	$1,30 \times 10^{-3}$	5,50	1,95-15,56
rs1426654	C, H	G	Cabelo Preto vs. Ruivo	$2,80 \times 10^{-3}$	21,83	2,89-164,90
rs16891982	A, B, C, D, E, F, H	C	Cabelo Preto vs. Loiro-Claro	$9,02 \times 10^{-6}$	33,72	7,13-159,40
rs10777129	H	A	Cabelo Preto vs. Loiro-Claro	$5,82 \times 10^{-3}$	7,95	1,82-34,67
rs4778232	E, F	T	Cabelo Preto vs. Loiro-Claro	$1,92 \times 10^{-4}$	5,60	2,27-13,87
rs1470608	C	T	Cabelo Preto vs. Loiro-Claro	$1,09 \times 10^{-4}$	7,49	2,70-20,77
rs1375164	F	T	Cabelo Preto vs. Loiro-Claro	$1,28 \times 10^{-4}$	5,85	2,37-14,43
rs4778138	F, G	G	Cabelo Preto vs. Loiro-Claro	$1,02 \times 10^{-4}$	12,38	3,48-44,07
rs4778241	E, F	C	Cabelo Preto vs. Loiro-Claro	$2,61 \times 10^{-2}$	6,02	1,24-29,24
rs16891982	A, B, C, D, E, F, H	C	Cabelo Preto vs. Castanho-Claro	$3,55 \times 10^{-16}$	7,39	4,57-11,95
rs28777	B, C, G	C	Cabelo Preto vs. Castanho-Claro	$2,88 \times 10^{-14}$	6,24	3,89-9,99
rs1448484	H	G	Cabelo Preto vs. Castanho-Claro	$6,21 \times 10^{-12}$	10,58	5,40-20,73
rs1470608	C	T	Cabelo Preto vs. Castanho-Claro	$6,65 \times 10^{-13}$	5,25	3,34-8,26
rs4778138	F, G	G	Cabelo Preto vs. Castanho-Claro	$1,06 \times 10^{-12}$	5,66	3,52-9,13
rs6497292	C	G	Cabelo Preto vs. Castanho-Claro	$2,29 \times 10^{-13}$	8,87	4,95-15,90
rs916977	F	T	Cabelo Preto vs. Castanho-Claro	$6,68 \times 10^{-14}$	5,56	3,55-8,711

Marcador	Sistema Preditivo*	Alelo	Característica	<i>p</i> exato de Fisher	<i>Odds Ratio</i>	IC 95%
rs1667394	C, E, F	C	Cabelo Preto vs. Castanho-Claro	$1,54 \times 10^{-13}$	5,55	3,52-8,75
rs1426654	C, H	G	Cabelo Preto vs. Castanho-Claro	$3,25 \times 10^{-17}$	15,69	8,28-29,73
rs6119471	C, H	G	Cabelo Preto vs. Castanho-Claro	$4,61 \times 10^{-6}$	11,23	3,99-31,62
rs1448484	H	G	Cabelo Preto vs. não Preto	$1,73 \times 10^{-17}$	6,38	4,16-9,78
rs1426654	C, H	G	Cabelo Preto vs. não Preto	$7,52 \times 10^{-22}$	5,92	4,119-8,51
rs1042602	B, C	A	Pele Clara vs. Escura	$4,24 \times 10^{-9}$	6,17	3,36-11,32
rs1126809	C	A	Pele Clara vs. Escura	$1,47 \times 10^{-5}$	5,46	2,53-11,77
rs17128291	C	G	Pele Clara vs. Escura	$8,83 \times 10^{-4}$	5,91	2,07-16,83
rs1800407	A, B, C, D, E, F	T	Pele Clara vs. Escura	$1,91 \times 10^{-2}$	11,04	1,48-82,33
rs1129038	C, D, E, F, G	T	Pele Clara vs. Escura	$2,99 \times 10^{-12}$	6,79	3,96-11,62
rs12913832	A, B, C, D, E, F, G	G	Pele Clara vs. Escura	$4,05 \times 10^{-12}$	6,60	3,87-11,24
rs11636232	E, F	T	Pele Clara vs. Escura	$7,07 \times 10^{-6}$	6,40	2,85-14,38
rs16891982	A, B, C, D, E, F, H	C	Pele Intermediária vs. Clara	$8,95 \times 10^{-18}$	5,17	3,55-7,52
rs1448484	H	G	Pele Intermediária vs. Clara	$9,06 \times 10^{-7}$	5,34	2,74-10,43
rs1426654	C, H	G	Pele Intermediária vs. Clara	$8,90 \times 10^{-15}$	7,33	4,43-12,12
rs6119471	C, H	G	Pele Intermediária vs. Clara	$1,12 \times 10^{-5}$	10,24	3,63-28,90
rs16891982	A, B, C, D, E, F, H	C	Pele Escura vs. Clara	$9,88 \times 10^{-18}$	14,37	7,81-26,42
rs28777	B, C, G	C	Pele Escura vs. Clara	$2,46 \times 10^{-16}$	9,34	5,47-15,93
rs10777129	H	A	Pele Escura vs. Clara	$5,66 \times 10^{-12}$	6,98	4,02-12,13
rs4778232	E, F	T	Pele Escura vs. Clara	$4,90 \times 10^{-13}$	6,33	3,84-10,45
rs1448484	H	G	Pele Escura vs. Clara	$6,86 \times 10^{-18}$	43,14	18,32-101,60
rs8024968	E, F	T	Pele Escura vs. Clara	$2,04 \times 10^{-11}$	5,86	3,50-9,84
rs1470608	C	T	Pele Escura vs. Clara	$6,22 \times 10^{-15}$	11,60	6,27-21,47
rs1375164	F	T	Pele Escura vs. Clara	$4,35 \times 10^{-14}$	7,76	4,56-13,21
rs4778138	F, G	G	Pele Escura vs. Clara	$8,40 \times 10^{-14}$	7,09	4,24-11,87
rs6497292	C	G	Pele Escura vs. Clara	$1,33 \times 10^{-12}$	5,94	3,63-9,71
rs916977	F	T	Pele Escura vs. Clara	$1,62 \times 10^{-13}$	5,27	3,39-8,20
rs1667394	C, E, F	C	Pele Escura vs. Clara	$5,54 \times 10^{-13}$	5,52	3,468-8,773
rs1426654	C, H	G	Pele Escura vs. Clara	$5,57 \times 10^{-17}$	30,96	13,86-69,14
rs6119471	C, H	G	Pele Escura vs. Clara	$5,04 \times 10^{-11}$	39,52	13,20-118,40
rs6058017	F, H	G	Pele Escura vs. Clara	$7,58 \times 10^{-12}$	5,67	3,45-9,32
rs16891982	A, B, C, D, E, F, H	C	Pele Escura vs. Intermediária	$5,47 \times 10^{-10}$	5,24	3,10-8,84
rs1448484	H	G	Pele Escura vs. Intermediária	$5,35 \times 10^{-14}$	7,78	4,56-13,29
rs1426654	C, H	G	Pele Escura vs. Intermediária	$1,40 \times 10^{-12}$	5,44	3,41-8,70
rs16891982	A, B, C, D, E, F, H	C	Pele Escura vs. não Escura	$1,07 \times 10^{-16}$	7,08	4,46-11,25
rs1448484	H	G	Pele Escura vs. não Escura	$1,51 \times 10^{-20}$	11,80	7,01-19,86
rs1470608	C	T	Pele Escura vs. não Escura	$1,93 \times 10^{-15}$	5,82	3,77-8,98
rs1375164	F	T	Pele Escura vs. não Escura	$2,08 \times 10^{-13}$	5,39	3,44-8,45

Marcador	Sistema Preditivo*	Alelo	Característica	<i>p</i> exato de Fisher	<i>Odds Ratio</i>	IC 95%
rs1426654	C, H	G	Pele Escura vs. não Escura	$4,61 \times 10^{-21}$	8,34	5,36-12,96
rs6119471	C, H	G	Pele Escura vs. não Escura	$1,46 \times 10^{-15}$	6,47	4,09-10,23

* A = *IrisPlex*; B = *HIrisPlex*; C = *HIrisPlex-S*; D = *Snipper7*; E = *Snipper13*; F = *Snipper23*; G = *Snipper12*; H = *Snipper10*.

APÊNDICE C

Tabela C1 – Lista do número de associações alélicas encontradas para cada um dos 61 marcadores analisados.

Marcador	Gene	Sem ajuste											
		Com ajuste de ancestralidade			Associações fracas			Associações moderadas			Associações fortes		
					(1 < OR < 5)			(5 ≤ OR < 10)			(OR ≥ 10)		
		Olhos	Cabelos	Pele	Olhos	Cabelos	Pele	Olhos	Cabelos	Pele	Olhos	Cabelos	Pele
rs16891982 ^{abc}	SLC45A2	-	-	-	3	4	1	-	1	3	-	3	1
rs28777 ^{bc}	SLC45A2	-	1	-	3	4	4	-	1	1	-	1	-
rs26722 ^a	SLC45A2	-	-	-	-	2	2	-	-	-	-	-	-
rs13289 ^c	SLC45A2	-	-	-	1	7	4	-	-	-	-	-	-
rs12203592 ^{abc}	IRF4	-	-	-	2	-	3	-	-	-	-	-	-
rs4959270 ^b	EXOC2	-	-	-	-	2	-	-	-	-	-	-	-
rs683 ^{abc}	TYRP1	-	-	-	3	4	3	-	-	-	-	-	-
rs1408799 ^{ac}	TYRP1	-	-	-	3	3	3	-	-	-	-	-	-
rs10756819 ^c	BNC2	-	-	-	2	6	4	-	-	-	-	-	-
rs1393350 ^{ac}	TYR	-	-	-	4	5	4	-	-	-	-	-	-
rs1042602 ^{bc}	TYR	-	-	-	2	4	3	-	-	1	-	-	-
rs1126809 ^c	TYR	-	-	-	4	5	3	-	-	1	-	-	-
rs35264875 ^b	TPCN2	-	-	-	2	3	2	-	1	-	-	-	-
rs3829241 ^c	TPCN2	-	-	-	4	5	3	-	-	-	-	-	-
rs12821256 ^{bc}	KITLG	-	-	-	-	-	-	-	-	-	-	-	-
rs10777129 ^c	KITLG	-	-	-	2	4	3	1	1	1	-	-	-
rs12896399 ^{ac}	SLC24A4	-	1	-	3	4	3	-	1	-	-	-	-
rs2402130 ^{bc}	SLC24A4	-	1	-	2	3	3	-	1	-	-	1	-
rs17128291 ^c	SLC24A4	-	-	-	-	-	2	-	-	1	-	-	-

Marcador	Gene	Sem ajuste											
		Com ajuste de ancestralidade			Associações fracas			Associações moderadas			Associações fortes		
					(1 < OR < 5)			(5 ≤ OR < 10)			(OR ≥ 10)		
		Olhos	Cabelos	Pele	Olhos	Cabelos	Pele	Olhos	Cabelos	Pele	Olhos	Cabelos	Pele
rs1426654 ^c	SLC24A5	-	-	-	2	2	2	1	1	2	-	2	1
rs12913832 ^{abc}	HERC2	-	1	-	1	3	3	1	2	1	3	5	-
rs2238289 ^c	HERC2	2	1	-	-	4	4	3	-	-	1	-	-
rs6497292 ^c	HERC2	-	2	-	-	4	3	2	1	1	1	-	-
rs1129038 ^{abc}	HERC2	-	2	1	1	3	3	1	2	1	3	5	-
rs1667394 ^{ac}	HERC2	-	-	-	1	5	3	2	1	1	1	-	-
rs11636232 ^a	HERC2	-	1	-	2	4	3	2	3	1	-	-	-
rs7183877 ^a	HERC2	1	-	-	1	1	-	1	-	-	-	-	-
rs8024968 ^a	HERC2	-	-	1	3	5	3	-	1	1	-	-	-
rs12592730 ^a	HERC2	-	1	-	2	4	3	-	1	-	1	-	-
rs916977 ^a	HERC2	-	1	-	1	5	3	2	1	1	1	-	-
rs1800407 ^{abc}	OCA2	-	-	-	-	1	1	-	-	-	-	-	1
rs1800414 ^c	OCA2	-	-	-	-	1	-	-	-	1	-	-	-
rs1470608 ^c	OCA2	-	-	1	3	5	2	-	2	1	-	-	1
rs1545397 ^c	OCA2	-	-	-	-	-	3	-	-	-	-	-	-
rs12441727 ^c	OCA2	-	-	-	2	2	2	-	-	-	-	-	-
rs4778232 ^a	OCA2	-	1	-	3	6	3	-	1	1	-	-	-
rs4778241 ^a	OCA2	-	1	-	1	4	3	2	3	-	-	-	-
rs1375164 ^a	OCA2	-	-	-	3	6	2	-	2	2	-	-	-
rs4778138 ^{ab}	OCA2	-	-	-	1	4	3	2	2	1	-	1	-
rs7495174 ^{ab}	OCA2	-	2	-	-	3	3	3	1	-	-	-	-
rs1448484 ^c	OCA2	-	-	-	-	2	-	2	1	2	1	2	2

Marcador	Gene	Sem ajuste											
		Com ajuste de ancestralidade			Associações fracas (1 < OR < 5)			Associações moderadas (5 ≤ OR < 10)			Associações fortes (OR ≥ 10)		
		Olhos	Cabelos	Pele	Olhos	Cabelos	Pele	Olhos	Cabelos	Pele	Olhos	Cabelos	Pele
rs885479 ^{bc}	MC1R	-	-	-	1	-	-	-	-	-	-	-	-
rs796296176 ^b	MC1R	-	-	-	-	-	-	-	-	-	-	1	-
rs1805005 ^b	MC1R	-	-	-	-	3	2	-	-	-	-	-	-
rs1805006 ^{bc}	MC1R	-	-	-	-	-	-	-	-	-	-	-	-
rs1805007 ^{bc}	MC1R	-	1	-	-	-	-	-	-	-	-	4	-
rs1805008 ^{bc}	MC1R	-	-	-	-	-	1	-	1	-	-	4	-
rs1805009 ^b	MC1R	-	-	-	1	-	2	-	1	-	-	-	-
rs2228479 ^{bc}	MC1R	-	-	-	-	-	-	-	-	-	-	-	-
rs11547464 ^{bc}	MC1R	-	-	-	-	-	-	-	-	-	-	4	-
rs1110400 ^{bc}	MC1R	-	-	-	-	-	-	-	-	-	-	-	-
rs201353883 ^b	MC1R	-	-	-	-	-	-	-	-	-	-	-	-
rs3212355 ^c	MC1R	-	-	-	-	-	-	-	-	-	-	-	-
rs3114908 ^c	ANKRD11	-	-	-	1	4	-	-	-	-	-	-	-
rs8051733 ^c	DEF8	-	1	-	-	4	-	-	4	-	-	1	-
rs12931267 ^b	FANCA	-	1	-	-	-	-	-	-	-	-	4	-
rs2378249 ^{bc}	ASIP	-	-	1	-	-	-	-	-	-	-	-	-
rs6119471 ^c	ASIP	-	-	-	2	2	1	-	1	1	-	1	1
rs1015362 ^a	ASIP	-	-	-	3	2	4	-	-	-	-	-	-
rs6058017 ^{ac}	ASIP	-	-	-	3	4	3	-	-	1	-	-	-
rs6059655 ^c	RALY	-	-	-	-	-	-	-	-	-	-	-	-

Obs: ^a Marcador utilizado para predição da cor dos olhos; ^b Marcador utilizado para predição da cor dos cabelos; ^c Marcador utilizado para predição da cor da pele.

Anexos

Anexo A – Termo de consentimento livre e esclarecido

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 n o _____,
 residente na _____ 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 _____ nº _____, 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.

1. EXTRAÇÃO DE DNA GENÔMICO	12020
2. QUANTIFICAÇÃO E VERIFICAÇÃO DA QUALIDADE DO DNA	12020
3. PREPARO DOS REAGENTES PARA ENRIQUECIMENTO DE ALVO	121
4. ENRIQUECIMENTO DO ALVO E PREPARO DA BIBLIOTECA	12121
4.1. Digestão e hibridação do DNA	122
4.2. Captura, ligação, eluição e PCR	12323
4.3. Purificação da biblioteca	12424
5. CONTROLE DE QUALIDADE E QUANTIFICAÇÃO DA BIBLIOTECA	124
6. SEQUENCIAMENTO	1255
7. REFERÊNCIAS	1266

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 ressuspenso 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 8: 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.

1. ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO	1288
1.1. Análise primária	1288
1.2. Análise secundária	1299
1.2.1. Avaliação da qualidade	1299
1.2.2. Trimagem	1299
1.2.3. Alinhamento ao genoma de referência	1311
1.2.3.1. SAM	1333
1.2.3.2. BAM	1344
1.3. Análise terciária	1355
1.3.1. Determinação de variantes	1355
1.3.1.1. VCF	1399
1.3.2. Remoção de genótipos de baixa qualidade	1411
2. REFERÊNCIAS	1433

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 *mem* 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:

```
PIG0001.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 \
--emiteRefConfidence 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., Kernytsky, 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>