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

“Caracterização da diversidade genética das regiões regulatórias e codificantes do gene *SLC24A5* e associações com fenótipos de pigmentação humana em amostra da população brasileira”.

Amanda Beatriz Candelária da Silva

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

RIBEIRÃO PRETO - SP

2019

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

Amanda Beatriz Candelária da Silva

Caracterização da diversidade genética das regiões regulatórias e codificantes do gene *SLC24A5* e associações com fenótipos de pigmentação humana em amostra 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 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.

Silva, Amanda Beatriz Candelária da

Caracterização da diversidade genética das regiões regulatórias e codificantes do gene *SLC24A5* e associações com fenótipos de pigmentação humana em amostra da população brasileira, 2019.

157 p. : il. ; 30cm

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

Orientador: Mendes-Junior, Celso Teixeira.

1. Genética de populações. 2. Fenotipagem Forense por DNA. 3. Pigmentação. 4. *SLC24A5*.

Nome: Silva, Amanda Beatriz Candelária da

Título: Caracterização da diversidade genética das regiões regulatórias e codificantes do gene *SLC24A5* e associações com fenótipos de pigmentação humana em amostra 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 minha mãe, meu maior exemplo de vida.

Agradecimentos

Aos meus pais, Andrea e Haroldo, que mesmo sem entender completamente este trabalho sempre acreditaram em mim, me apoiaram e estiveram ao meu lado, apesar dos 300 km de distância.

A minha avó Rosalina (*in memorian*) que esteve ao meu lado sonhando e comemorando cada uma das minhas conquistas, que se preocupava excessivamente com cada dor de garganta e que sempre fez o possível pela felicidade das filhas e netos.

Ao meu namorado, Bruno, que sempre esteve comigo para me acalmar quando tudo dava errado e compartilhar a felicidade quando tudo dava certo. Agradeço imensamente a amizade, o companheirismo, o cuidado e amor que temos um pelo outro. Você me faz melhor a cada dia.

Aos amigos que cultivei desde os primeiros meses da graduação, todos que amam o OVATAB, em especial Bárbara, Luandra e Maike, que estiveram ao meu lado em momentos difíceis para me consolar e em momentos felizes para comemorar. Vocês se tornaram uma verdadeira família para mim.

Aos meus amigos Guilherme e Denis que fizeram meus dias de mestranda mais engraçados, leves e com mais sorvete, chocolate e bolos.

Aos meus queridos amigos do Laboratório de Pesquisas Forenses e Genômica: Alison, Malu, Guilherme (IC), Nádia, Guilherme, Thássia, Letícia e Tamara. Vocês me acolheram e me auxiliaram em tudo o que podiam, sem vocês não teria chego ao fim deste trabalho, todos merecem muito crédito.

Ao meu orientador, Prof. Dr. Celso, por me acolher em seu grupo de pesquisa tão bem selecionado, pelas toneladas de paciência e por me guiar até aqui. Durante estes dois anos de mestrado tive a oportunidade de evoluir como pessoa e profissional.

Ao Prof. Dr. Aguinaldo Luiz Simões por ceder o uso de equipamentos de seu laboratório.

Ao Prof. Dr. Eduardo Antônio Donadi e ao Laboratório de Imunopatologia Molecular pela estrutura cedida.

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

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) e à estrutura cedida pela Universidade de São Paulo que tornaram esta pesquisa possível.

Aos membros da banca examinadora por compartilharem seus conhecimentos e contribuir com este trabalho e com meu crescimento profissional.

Aos voluntários que se disponibilizaram e confiaram neste trabalho, compartilhando parte de si e de sua história.

Muito obrigada a todos!

Resumo

SILVA, A. B. C. **Caracterização da diversidade genética das regiões regulatórias e codificantes do gene *SLC24A5* e associações com fenótipos de pigmentação humana em amostra da população brasileira.** 2019. 157 páginas. Dissertação de Mestrado – Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brasil.

A Fenotipagem Forense por DNA é uma metodologia que busca identificar características externas visíveis de indivíduo desconhecido a partir do material biológico encontrado em uma cena de crime, sendo possível afunilar a lista de suspeitos em uma investigação. O fenótipo mais visível e estudado para este fim são os fenótipos de pigmentação, ou seja, a cor da pele, dos olhos e dos cabelos. Assim, torna-se necessário o estudo de genes envolvidos na via de produção de melanina, a melanogênese, a fim de identificar marcadores genéticos que possam ser empregados para fins de predição. O gene *SLC24A5* está envolvido neste processo pela produção da proteína NCKX5. Esta proteína encontra-se na membrana de melanosomos e atua como um trocador iônico de sódio e cálcio dependente de potássio. Desta forma, é possível controlar o pH no interior desta célula, o que interfere no tipo e quantidade de melanina produzida, feomelanina (que gera tonalidade amarela a vermelha) ou eumelanina (que gera tonalidade marrom a preta). Adicionalmente, este gene possui mutações relacionadas ao albinismo oculocutâneo 6 (OCA6), causador de hipopigmentação na pele, nos olhos e no cabelo, podendo causar também estrabismo e fotofobia. Este trabalho buscou encontrar sítios de variação presentes em uma população amostrada na região de Ribeirão Preto que estejam associados a fenótipos de pigmentação, procurando entender ainda a causa de tal associação a partir da interação com Fatores de Transcrição (TFs) e miRNAs, bem como pela alteração da sequência primária da proteína traduzida. Foram encontrados 65 sítios de variação na amostra populacional, 13 dos quais tiveram seus alelos ou genótipos associados a algum fenótipo de pigmentação. Dentre eles, apenas um está presente na região codificante, o SNP rs1426654 já extensamente associado a fenótipos de pigmentação e conhecido como Ala111Thr, cuja mutação causa alteração na proteína produzida. Além disso, foram encontrados quatro sítios de variação na região promotora associados a fenótipos de pigmentação. Para três deles foi possível encontrar TFs com maior afinidade pelo alelo associado a fenótipos escuros do sítio em

questão. Em relação aos haplótipos, foram encontrados 39 na região de Ribeirão Preto e sete foram associados a fenótipos de pigmentação que confirmam as associações encontradas para alelos e genótipos. Os resultados deste estudo demonstram a importância deste gene e de seus sítios de variação, particularmente na região promotora, para estudos de fenótipos de pigmentação.

Palavras-chave: genética forense; Fenotipagem Forense por DNA; pigmentação; *SLC24A5*.

Abstract

SILVA, A. B. C. Characterization of the genetic diversity of regulatory and coding regions of the *SLC24A5* gene and association with human pigmentation phenotypes in a Brazilian population sample. 2019. 157 pages. Dissertation (Master's Degree) - Faculty of Philosophy, Sciences and Letters at Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brasil.

Forensic DNA Phenotyping is a methodology that seeks to identify visible external characteristics of an unknown individual from the biological material found in a crime scene, making it possible to reduce the list of suspects in an investigation. The most visible and studied phenotypes for this purpose are pigmentation phenotypes, i.e., the skin, eyes and hair color. Thus, it is necessary to study the genes involved in the melanin production pathway, the melonogenesis, in order to identify genetic markers useful for predictive purposes. The *SLC24A5* gene is involved in this process by the production of the protein NCKX5. This protein is found in the membrane of melanosomes and acts as a sodium and calcium ion exchanger dependent on potassium. In this way, it is possible to control the pH inside this cell, which interferes in the type and quantity of melanin produced, pheomelanin (which generates yellow to red pigment) or eumelanin (which generates brown to black pigment). In addition, this gene has mutations related to oculocutaneous albinism 6 (OCA6), causing hypopigmentation in the skin, eyes and hair, and may also cause strabismus and photophobia. This study sought to find variation sites present in a population sampled in the region of Ribeirão Preto that are associated with pigmentation phenotypes, trying to understand the cause of such associations from the interaction with Transcription Factors and miRNAs, as well as from changes in the primary sequence of the translated protein. There were 65 variation sites in the obtained sample, 13 of which had their alleles or genotypes associated with some pigmentation phenotype. Among them, only one is present in the coding region, the rs1426654 SNP already extensively associated with pigmentation phenotypes and known as Ala111Thr, whose mutation causes alteration in the translated protein. In addition, four variation sites were found in the promoter region associated with pigmentation phenotypes. It was possible to find TFs with greater affinity for the allele associated with dark phenotypes from three of these variation sites. Regarding haplotypes, 39 were found in the Ribeirão Preto region and seven were associated with

pigmentation phenotypes that confirm the associations found for alleles and genotypes. The results of this study demonstrate the importance of this gene and its variation sites, particularly the promoter region, for studies of pigmentation phenotypes.

Keywords: forensic genetics; Forensic DNA Phenotyping; pigmentation; *SLC24A5*.

Lista de Figuras

- Figura 1** – Processo de melanogênese a partir de tirosina na ausência e na presença de cisteína, produzindo eumelanina e feomelanina, respectivamente.....20
- Figura 2** – Representação da localização de NCKX5 no melanossomo, proteína transmembrana melanossomal para transporte de íons Ca^{2+} , Na^{+} e K^{+} , com implicação decisiva no processo de produção de melanina.....22
- Figura 3** – Representação da estrutura do gene *SLC24A5* destacando o seu SNP mais conhecido, rs1426654, onde ocorre a troca de uma guanina por uma adenina no códon 111, alterando o aminoácido traduzido de alanina para treonina.....23
- Figura 4** – Captura de tela obtida no *Integrative Genomics Viewer 2.3.75* demonstrando as porções do gene *SLC24A5* cobertas pelo kit *HaloPlex Target Enrichment System* (Agilent Technologies, Inc.). A figura é dividida em dois quadrantes: o primeiro apresenta o cromossomo 15, onde a barra vermelha indica a posição do gene *SLC24A5*; o segundo quadrante apresenta em verde as regiões cobertas, os blocos azuis que representam os éxons deste gene e as linhas azuis que representam seus íntrons.....37
- Figura 5** – Características da amostra populacional obtida em Ribeirão Preto de acordo com os fenótipos de pigmentação da pele, dos olhos e dos cabelos e a ancestralidade dos indivíduos segundo o IBGE.....45
- Figura 6** – Distribuição bidimensional das distâncias obtidas entre Ribeirão Preto e as populações avaliadas pelo Projeto *1000 Genomes*. A distribuição foi realizada de acordo com os valores de distância genética (F_{ST}) calculados.....60
- Figura 7** – Distribuição bidimensional das distâncias obtidas entre Ribeirão Preto e os grupos populacionais avaliados pelo Projeto *1000 Genomes*. A distribuição foi realizada de acordo com os valores de distância genética (F_{ST}) calculados.....60
- Figura 8** – Representação do Desequilíbrio de Ligação (DL) observado no gene *SLC24A5* obtido de acordo com o método de Gabriel, et al. (2002). Os quadrados vermelhos indicam intenso DL, LOD (Log of Odds) ≥ 2 e $D' = 1$, os brancos indicam ausência de DL, $LOD < 2$ e $D' < 1$, e os azuis indicam fraco, porém completo DL, $LOD < 2$ e $D' = 1$. Encontramos um grande bloco de aproximadamente 14 kb em forte DL devido aos SNPs rs2675346 (7), rs2433354 (13), rs2433356 (24) e rs142665 (27). No canto inferior direito encontram-se as frequências dos haplótipos levando em conta os quatro sítios de variação.....61
- Figura 9** – Características de fragmento da região promotora do gene *SLC24A5*, onde se encontram os sítios de variação rs553733344, rs55728404 e rs2675346 (destacados em azul claro), alvos de Fatores de Transcrição (TFs) e associados às marcas de histonas H3K4me1_Enh, H3K4me3_Pro e H3K27ac_Enh.....70
- Figura 10** – Características de fragmento que compõe a região 3'UTR do gene *SLC24A5*, onde se encontram 18 sítios de variação.....86

Lista de Tabelas

Tabela 1 – Populações avaliadas pelo Projeto <i>1000 Genomes</i> (Fase 3) distribuídas conforme os cinco grupos populacionais amostrados.....	35
Tabela 2 – Dados de número de <i>reads</i> , aproveitamento e qualidade obtidos nos sete ensaios de sequenciamento realizados e a média de cada parâmetro.....	46
Tabela 3 – Sítios de variação (excluindo-se <i>singletons</i>) encontrados na amostra populacional de Ribeirão Preto, suas posições no cromossomo 15 utilizando o genoma hg19 como referência e regiões em que se encontram no gene <i>SLC24A5</i>	47
Tabela 4 – Frequências de alelos alternativos e referências dos 31 sítios de variação avaliados, seus valores de heterozigose observada e esperada e aderência ao Equilíbrio de Hardy-Weinberg para a população de Ribeirão Preto. Em negrito estão os sítios de variação polimórficos. Sublinhados estão os sítios de variação que desviam do Equilíbrio de Hardy-Weinberg.....	49
Tabela 5 – Haplótipos do gene <i>SLC24A5</i> encontrados na região de Ribeirão Preto indicando o alelo que possui em cada sítio de variação identificado. Os pontos observados nas diferentes posições indicam que o haplótipo em questão possui a base presentes na sequência de referência.....	51
Tabela 6 – Frequências dos haplótipos encontrados na região de Ribeirão Preto e nos grupos populacionais avaliados pelo Projeto <i>1000 Genomes</i>	54
Tabela 7 – Diversidade haplotípica do gene <i>SLC24A5</i> calculada para a amostra populacional da região de Ribeirão Preto e para os grupos populacionais estudados pelo Projeto <i>1000 Genomes</i>	56
Tabela 8 – Índices de F_{ST} baseado em frequências haplotípicas estimados entre a amostra populacional de Ribeirão Preto e as populações amostradas pelo Projeto <i>1000 Genomes</i>	57
Tabela 9 – Teste exato de diferenciação baseado em frequências haplotípicas estimado entre a amostra populacional de Ribeirão Preto e populações amostradas pelo Projeto <i>1000 Genomes</i>	58
Tabela 10 – Índices de F_{ST} baseado em frequências haplotípicas estimados entre a amostra populacional de Ribeirão Preto e os grupos populacionais amostrados pelo Projeto <i>1000 Genomes</i>	59
Tabela 11 – Associações entre alelos ou genótipos e fenótipos de pigmentação que demonstraram menores valores de probabilidade por meio do teste exato de Fisher. Estão aqui apresentadas apenas as associações com $p < 1,000 \times 10^{-15}$. O conjunto completo de associações obtidas pode ser encontrado no Apêndice A.....	63
Tabela 12 – Associações estatisticamente significantes entre sete haplótipos e fenótipos de pigmentação.....	65
Tabela 13 – Fatores de Transcrição (TFs) cuja eficiência de ligação ao gene <i>SLC24A5</i> pode ser alterada pela presença de alelo referência ou alternativo nos sítios de variação	

associados a algum fenótipo de pigmentação. A afinidade do TF por cada alelo é avaliada pelo valor de Δ LOD.....68

Tabela 14 – Conjunto de 20 miRNAs que apresentam os maiores valores de estabilidade na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*.....72

Tabela 15 – Conjunto de 20 miRNAs que apresentam os maiores valores de especificidade na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*.....72

Tabela 16 – Conjunto de 20 miRNAs que apresentam os maiores valores de score combinado de estabilidade e especificidade na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*.....73

Tabela 17 – Conjunto de 32 miRNAs que apresentam alta estabilidade, especificidade e/ou *score* combinado na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*. Em negrito estão os miRNAs que apresentam altos valores em todos os três parâmetros avaliados.....74

Tabela 18 – Médias do índice de melanina relacionadas a genótipos de sítios de variação associados a fenótipos de pigmentação. Valores significativos de *p* para os testes de Mann-Whitney e Kruskal-Wallis. Em negrito estão os valores de *p* > 0,05, ou seja, não significativo.....76

Sumário

1. Introdução.....	16
1.1. Fenotipagem Forense por DNA.....	17
1.2. Melanogênese	19
1.3. <i>SLC24A5</i>	22
1.4. Miscigenação da população brasileira	24
2. Justificativa.....	26
3. Hipótese.....	28
4. Objetivos.....	30
4.1. Objetivos gerais	31
4.2. Objetivos específicos	31
5. Material e Métodos.....	32
5.1. Obtenção da amostra populacional de Ribeirão Preto (Anexo A).....	33
5.2. Obtenção de outras amostras populacionais (Projeto <i>1000 Genomes</i>)	34
5.3. Procedimentos experimentais	36
5.3.1. Extração (Anexo B – item 1)	36
5.3.2. Quantificação e verificação de qualidade do DNA genômico (Anexo B – item 2).....	36
5.3.3. Preparo e validação de bibliotecas (Anexo B – itens 3 a 6)	36
5.3.4. Sequenciamento (Anexo B – item 7).....	38
5.4. Procedimentos computacionais (Anexo C).....	39
5.4.1. Trimagem e mapeamento.....	39
5.4.2. Obtenção de genótipos.....	39
5.4.3. Obtenção de haplótipos (Anexo D – item 2)	40
5.4.4. Análises estatísticas (Anexos D e E)	41
6. Resultados.....	43
6.1. Amostra populacional	44
6.2. Sequenciamento	45
6.3. Variant calling.....	46
6.4. Frequências alélicas, heterozigose, frequências haplotípicas, e diversidade haplotípica.....	48
6.5. Diferenciação populacional	56
6.6. Desequilíbrio de Ligação	61
6.7. Associações fenotípicas	62
6.8. Interações TF/TFBS na região promotora do gene <i>SLC24A5</i>	66
6.9. Alterações na sequência primária da proteína NCKX5	71
6.10. Interações miRNA/mRNA na região 3'UTR do gene <i>SLC24A5</i>	71

6.11. Relação entre índice de melanina e fenótipos e genótipos	75
7. Discussão	78
7.1. Amostra populacional	79
7.2. Desvios do equilíbrio de Hardy-Weinberg e Desequilíbrio de Ligação	79
7.3. Associações fenotípicas	80
7.3.1. Região Promotora	82
7.3.2. Região CDS	83
7.3.3. Região 3'UTR.....	85
8. Conclusões.....	87
Referências	90
Apêndices.....	99
Apêndice A – Tabela 11 (completa) – Associações entre alelos ou genótipos e fenótipos de pigmentação entre as 285 associações envolvendo 13 sítios de variação.	100
Apêndice B – Frequências globais dos alelos referência e alternativo dos sítios de variação que apresentaram associações com fenótipos na amostra populacional de Ribeirão Preto explicadas pela interação com Fatores de Transcrição ou alteração na sequência primária da proteína NCKX5.....	111
Anexos.....	112
Anexo A – Termo de Consentimento Livre e Esclarecido.....	113
Anexo B – Protocolo de análise laboratorial (HaloPlex) para sequenciamento de nova geração.....	116
Anexo C – Protocolo de processamento computacional de dados de sequenciamento de nova geração.....	124
Anexo D – Protocolo de análise estatística de dados de sequenciamento de nova geração.....	142
Anexo E – Protocolo para identificação computacional de interações entre miRNAs e região 3'UTR.....	152

Introdução

1. Introdução

1.1.Fenotipagem Forense por DNA

O DNA como ferramenta forense tem se mostrado um grande aliado para a solução de crimes, assim como a papiloscopia e a odontologia forense. Entretanto, todas essas metodologias requerem uma amostra referência que possibilite determinar a origem de uma amostra desconhecida coletada em cena de crime, o que se torna um obstáculo em muitas situações.

No caso do DNA, este obstáculo pode ser superado pela existência de suspeitos ou por bancos de DNA, que têm auxiliado o processo de identificação forense em vários países. No entanto, esta ferramenta pode não ser de grande utilidade se a amostra desconhecida não for compatível com uma amostra previamente inserida no banco. No Brasil, este empecilho é ainda mais evidente, pois o Banco Nacional de Perfis Genéticos, alimentado pelos laboratórios que compõem a Rede Integrada de Bancos de Perfis Genéticos (RIBPG), ainda é pequeno, contendo uma parcela não significativa da população, com apenas 30.809 perfis de DNA distribuídos entre vestígios encontrados em cenas de crime, indivíduos condenados, suspeitos, restos mortais não identificados e de identidade desconhecida e referências diretas e indiretas de pessoas desaparecidas (RIBPG, 2019).

Percebemos, portanto a necessidade de metodologias que permitam inferir a origem de uma amostra desconhecida sem uma amostra referência, cuja disponibilidade não é observada em diversos cenários. A Fenotipagem Forense por DNA (FDP, do inglês, *Forensic DNA Phenotyping*) tem se mostrado uma alternativa promissora para prever as características externas visíveis do indivíduo que originou a amostra encontrada em cena de crime, incluindo características como a cor da pele, dos olhos e do cabelo (Liu, et al. 2010a; Sulem, et al. 2007), morfologia do cabelo (Pospiech, et al. 2018; Pospiech, et al. 2015) e a altura (Lango Allen, et al. 2010), permitindo a obtenção de um “retrato falado molecular”, que poderá aproximar a investigação de possíveis suspeitos.

Um sistema amplamente conhecido e desenvolvido para fins forenses, o IrisPlex, foi criado para predição da cor dos olhos utilizando seis SNPs (Walsh, et al. 2011; Walsh, et al. 2012). Mais tarde, foram adicionados 18 SNPs para elaborar o sistema HIrisPlex, para a predição da cor dos cabelos e dos olhos (Walsh, et al. 2013). Em 2018, o sistema HIrisPlex-S foi desenvolvido com a adição de 17 SNPs para a predição da cor dos olhos, cabelos e pele (Chaitanya, et al. 2018).

Entretanto, o sistema HIrisPlex, foi validado apenas em populações não miscigenadas, europeias e asiáticas, e por isso não apresenta boa acurácia na predição de fenótipos em populações miscigenadas, como a população brasileira, onde há maior frequência de fenótipos intermediários. Um estudo realizado por nosso grupo de pesquisa com 340 indivíduos demonstrou que apenas 78,53% deles tiveram a cor de seus olhos predita corretamente e apenas 64,16% tiveram a cor de seus cabelos predita corretamente (Carratto 2017).

O uso de SNPs em fenotipagem é mais comum e eficiente por uma série de razões. Uma delas é a menor dificuldade de genotipagem em casos de amostras degradadas, uma grande preocupação quando tratamos de amostras forenses, já que SNPs, por terem variação única, são menos propensos a danos quando comparados a STRs. Outro motivo é que a variação de SNPs em regiões codificantes pode afetar a proteína traduzida, afetando a sua eficiência. Quando esta proteína possui papel na produção de características externas, como a pigmentação, podemos explicar a associação deste sítio de variação com o fenótipo (Cartegni, et al. 2002).

Contudo, o uso de SNPs para identificação forense se baseia em técnica mais trabalhosa quando comparado a STRs, marcadores mais frequentemente empregados devido à maior informatividade que oferecem (Musgrave-Brown, et al. 2007). Entretanto, vivemos um grande avanço em metodologias de sequenciamento que diminuem drasticamente os custos. O advento do sequenciamento de nova geração (NGS, do inglês *Next Generation Sequencing*) contribuiu significativamente para o decaimento de custos, pois permite o sequenciamento simultâneo de várias regiões do genoma e de vários indivíduos, ou seja, sequenciamento multiplex.

Com isso, os custos de sequenciamento reduziram de cerca de bilhões de dólares, quando o primeiro genoma humano foi sequenciado, para cerca de centenas de dólares. Entretanto, esta rápida evolução tecnológica, gerando um grande volume de

dados gerados, requer o emprego de inúmeras ferramentas de bioinformática e estatística para a interpretação dos resultados.

1.2.Melanogênese

O processo de melanogênese começa ainda na formação do embrião, a embriogênese. Durante esta fase os melanoblastos migram da crista neural para os locais onde se depositarão e diferenciarão em melanócitos, ou seja, epiderme, íris e folículos capilares. A pigmentação apresentada nestes locais dependerá, portanto, da quantidade e tipo de melanina ali produzida durante a vida do indivíduo (Hearing 2011; Kawakami e Fisher 2011; Liu, et al. 2013). Na pele, a cor observada também é brevemente influenciada pela quantidade de hemoglobina e carotenoides (Lin e Fisher 2007).

Este processo inicia-se com o stress oxidativo causado pela radiação ultravioleta, que causa um acúmulo de espécies reativas de oxigênio nos queratinócitos, ativando o gene *POMC* (*Proopiomelanocortina*). O produto proteico deste gene é clivado em peptídeos menores, produzindo α -MSH (α -hormônio estimulante de melanócito), que, por sua vez, ativa o receptor MC1R (*receptor de melanocortina-1*). A ativação de MC1R é responsável pelo aumento da produção de cAMP (monofosfato de adenosina cíclico), que culmina na ativação de *MITF* (*fator de transcrição associado a microftalmia*). Este fator de transcrição é responsável por regular a expressão de inúmeros genes com papel na via de melanogênese, como *TYR*, *TYRP1* e *DCT* (Kim, et al. 2017).

A síntese de melanina ocorre em vesículas denominadas melanossomos com a participação do complexo de Golgi, ambos encontrados em células especializadas chamadas de melanócitos. Este processo ocorre por meio de reações de oxidação da tirosina (Figura 1), se dividindo posteriormente em duas cadeias reacionais em função da presença dos componentes necessários para a produção de cada tipo de melanina, feomelanina e eumelanina.

A tirosina presente no melanossomo é inicialmente oxidada pela enzima tirosinase (TYR) a L-DOPA, que é novamente oxidada pela tirosinase a DOPAquinona. Neste momento, ocorre uma bifurcação no processo. Na presença de cisteína é formada

a cisteinil-DOPA, que é posteriormente oxidada e polimerizada para produzir a feomelanina, o pigmento claro responsável por gerar as cores amarela e vermelha. Já na ausência de cisteína a DOPAquinona espontaneamente se transforma em DOPAcromo. Esta molécula, na presença da enzima DCT, também conhecida como TYRP2, é oxidada a DHICA (ácido-5,6-diidroxiindol-2-carboxílico) e, em seguida, oxidada por TYRP1 a DHICA-melanina. O DOPAcromo, espontaneamente também se oxida, formando DHI (5,6-dihidroxiindol), que, por sua vez é oxidada a DHI-melanina. A DHICA-melanina é o tipo mais claro de eumelanina, responsável pela cor marrom, enquanto a DHI-melanina é o tipo mais escuro de eumelanina, responsável pela cor preta (Cichorek, et al. 2013).

Figura 1 – Processo de melanogênese a partir de tirosina na ausência e na presença de cisteína, produzindo eumelanina e feomelanina, respectivamente (Cichorek, et al. 2013).

Tratando-se de um processo oxidativo, o pH do melanócito é um fator de grande importância na melanogênese, podendo alterar o tipo e a quantidade de melanina produzida. Sabe-se que o pH ótimo da tirosinase é igual a 6,8, o que significa que um

pH básico ou muito ácido acarretam em danos a melanogênese, causando transtornos de hipopigmentação devido à produção ineficiente de melanina (Ancans, et al. 2001).

A melanina produzida ao final do processo de melanogênese é transportada para os queratinócitos adjacentes da pele ou para os folículos capilares e assim podem se manifestar na forma de pigmentação visível. Na íris, os melanossomos são retidos pelos melanócitos na úvea (estrutura formada pelo corpo ciliar, íris e coróide) e assim a pigmentação dos olhos se manifesta (Sturm, et al. 1998).

Apesar de terem a mesma origem embrionária, o tipo e a quantidade de melanina em diferentes tecidos podem variar em função da ativação independente de genes que interferem no processo de melanogênese. São estes fatores que nos permite observar fenótipos como pele clara, cabelos escuros e olhos azuis (Sturm, et al. 2001).

A presença e tipos de melanina na pele foram selecionados de acordo com a incidência natural de radiação UV no ambiente dos indivíduos. Assim populações com peles claras devido à maior quantidade de feomelanina, como europeus e asiáticos, frequentemente residentes em locais com menor incidência de luz, têm absorção mais eficiente, porém menor proteção solar. Enquanto isso, populações com peles mais escuras, devido a maiores quantidades de eumelanina, como africanos e sul asiáticos, frequentemente residentes em locais onde a incidência de luz é alta, possuem menor eficiência na absorção de luz, porém maior proteção solar (Canfield, et al. 2013; Jablonski e Chaplin 2010).

A importância da pigmentação na absorção de luz está relacionada à produção de vitamina D, um processo que requer a exposição do indivíduo a radiação UV. Esta vitamina é responsável pela absorção de cálcio pelos ossos e sua ausência pode acarretar em fraquezas e más-formações ósseas, como osteoporose e raquitismo (Rossberg, et al. 2016). Dessa forma, indivíduos de peles claras tendem a produzir maiores quantidades de vitamina D, devido à absorção de luz facilitada pela cor da pele (Bonilla, et al. 2014).

No que se refere à proteção solar, as moléculas de melanina podem agir como protetores solares devido à alta conjugação de suas estruturas, diminuindo a absorção de radiação UV pela pele (Maddodi, et al. 2012). Essa proteção garante maior defesa contra este tipo de radiação aos indivíduos de peles mais escuras, impedindo o

desenvolvimento de mutações que possam causar doenças relacionadas à intensa exposição solar, como o melanoma (Henriksen, et al. 1990).

1.3. *SLC24A5*

O gene *SLC24A5* (do inglês, *solute carrier family 24 member 5*) está localizado no braço longo do cromossomo 15 (15q21.1), contém mais de 21kb distribuídas em nove éxons que codificam 500 aminoácidos, transcrevendo a proteína transmembrana melanossomal NCKX5 (Figura 2). Mutações neste gene estão associadas a um dos sete tipos de albinismo oculocutâneo descobertos até o momento, o OCA6 (Bertolotti, et al. 2016; Morice-Picard, et al. 2014), denominação também usada, mais recentemente, em referência ao gene.

Figura 2 - Representação da localização de NCKX5 no melanossomo, proteína transmembrana melanossomal para transporte de íons Ca^{2+} , Na^{+} e K^{+} , com implicação decisiva no processo de produção de melanina (Ainger, et al. 2017).

A família de proteínas NCKX abrange cinco proteínas transmembranas responsáveis pela regulação de sódio e cálcio dependente de potássio, trocando uma molécula de Ca^{2+} por quatro moléculas de Na^{+} e uma molécula de K^{+} . Contudo as funções de cada proteína NCKX ainda são pouco detalhadas. Sabe-se que NCKX1 atua no transporte de cálcio em células da retina (Schnetkamp, et al. 1991). NCKX2 é encontrada no cérebro e em células fotorreceptoras do olho (Li, et al. 2006; Prinsen, et al. 2000). NCKX3 também é encontrada no cérebro e em músculos lisos (Kraev, et al. 2001). NCKX4 é encontrada no coração, estômago, rins e, principalmente, no

hipocampo (Li, et al. 2002; Li e Lytton 2014; Stephan, et al. 2011). NCKX5 está localizada no complexo de Golgi e melanossomos de melanócitos, possuindo papel na melanogênese (Ginger, et al. 2008). Além de regular a troca iônica entre o melanossomo e o endossomo a proteína NCKX5 também possui o papel de regular o pH melanossomal, o que interfere na maturação do melanossomo e no tipo de melanina produzido (Lamason, et al. 2005).

O SNP mais conhecido do gene *SLC24A5* é o rs1426654 (Figura 3), responsável pelas primeiras associações do gene à ancestralidade e pigmentação. Este SNP é frequentemente denominado Ala111Thr, pois a presença do alelo A no códon que codifica o 111º aminoácido leva à tradução de treonina, enquanto o alelo ancestral G nesta posição leva à tradução do aminoácido alanina. Inúmeros estudos demonstram associação do alelo A, com frequência de 98% a 100% na população europeia, a peles claras e do alelo G, com frequência de 93% a 100% nas populações africana, ameríndia e do leste asiático, a peles escuras (Lamason, et al. 2005).

Figura 3 – Representação da estrutura do gene *SLC24A5* destacando o seu SNP mais conhecido, rs1426654, onde ocorre a troca de uma guanina por uma adenina no códon 111, alterando o aminoácido traduzido de alanina para treonina (Basu Mallick, et al. 2013).

Outros poucos SNPs deste gene foram avaliados em estudos de associação com fenótipos de pigmentação, de identificação de Marcadores Informativos de Ancestralidade (AIM, do inglês *Ancestry Informative Markers*) ou de associação com câncer de pele. Os sítios de variação rs16960620 e rs2555364, localizados no íntron 2, foram estudados com rs1426654 formando haplótipos propostos como AIMs, sendo uma alternativa ao rs1426654 (Giardina, et al. 2008a). O SNP rs2470102, localizado no íntron 8, foi identificado como candidato à associação com fenótipos de pigmentação, pois está em região de interação com miRNAs (Richardson, et al. 2011). Esta

associação foi confirmada utilizando dados de *DNA-pooling* para melanoma (Liu, et al. 2010b). O sítio rs17426596, localizado no íntron 2, foi avaliado em estudo de associação com risco de câncer de pele e fenótipos de pigmentação da pele e dos cabelos em indivíduos caucasianos, não sendo observadas quaisquer associações significativas com os fenótipos avaliados (Nan, et al. 2009). O sítio rs2675346, localizado na região promotora, se mostrou associado a variantes do gene *VDR* (Receptor de Vitamina D) utilizando dados de Desequilíbrio de Ligação (Tiosano, et al. 2016), o que poderia indicar associações com a pigmentação, visto a relação entre a absorção de luz solar com a produção de vitamina D. Observa-se, portanto, que pouco se sabe sobre a variabilidade genética deste gene.

A supressão do gene *SLC24A5* diminui significativamente os níveis de pigmentação, como demonstrado por estudos de *knockdown* (Ginger, et al. 2008) e associações com OCA6 (Bertolotti, et al. 2016; Morice-Picard, et al. 2014). Há relação também com a diminuição dos níveis das proteínas TYRP1 e TYR (Ginger, et al. 2008), e associação entre a proteína NCKX5 e os níveis de expressão de *MC1R* (Wilson, et al. 2013), todas proteínas de extrema importância na melanogênese.

Assim, o gene *SLC24A5* é um dentre os centenas de genes considerados de grande importância no processo de biossíntese de melanina e o conhecimento sobre seus SNPs e funções tem muito a contribuir tanto para a área forense quanto para a medicina, auxiliando o tratamento de doenças com sintomas de hipo ou hiperpigmentação, como OCA6.

1.4.Miscigenação da população brasileira

A população brasileira é uma população ideal para estudos de pigmentação devido ao seu amplo histórico de miscigenações e diversidade fenotípica. Em quase 520 anos de história, inúmeros movimentos migratórios ocorreram e produziram uma população extremamente heterogênea, composta principalmente por indivíduos de origem ameríndia, europeia e africana (Pena, et al. 2011).

Com o descobrimento do país, o primeiro grande fluxo migratório ocorreu trazendo principalmente portugueses para apropriação e exploração de terras, minerais e

flora. A população indígena sofreu então uma drástica redução devido às inúmeras doenças trazidas por europeus e aos conflitos entre tribos e colonizadores. Poucos anos mais tarde, um segundo fluxo migratório se iniciou com a chegada de africanos para trabalho escravo, durando mais de três séculos. Já no século XIX, um projeto de colonização agrícola e povoamento atraiu principalmente alemães e italianos para o país (Patarra e Fernandes 2011).

Este fluxo migratório ocorreu em diferentes níveis em cada região do país. Nos estados do Nordeste brasileiro, por exemplo, o fluxo migratório de africanos foi mais intenso, enquanto nos estados da região Sul do país o fluxo migratório europeu foi mais evidente (Pena, et al. 2011).

Esse fato é ainda observado pelo censo realizado a cada década no país pelo Instituto Brasileiro de Geografia e Estatística (IBGE) que leva em conta a ancestralidade dos entrevistados: em todo o Brasil 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; no Nordeste apenas 29,44% consideraram-se brancos e 9,53% dos entrevistados consideraram-se pretos; já na região Sul do país 78,47% consideraram-se brancos e apenas 4,05% consideraram-se pretos.

No estado de São Paulo, o censo (IBGE) revelou 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. Na mesorregião de Ribeirão Preto, 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.

Esta grande miscigenação torna a população brasileira ideal para estudos de fenótipos de pigmentação, pois é possível observar uma ampla variedade de fenótipos, que abrangem cores claras, escuras e especialmente intermediárias, que não são observadas com facilidade em populações não miscigenadas (Adhikari, et al. 2016; Lock-Andersen, et al. 1998; Park, et al. 2012). Isso permite a identificação de novos sítios de variação associados à pigmentação, que podem ser aplicados em atividades de fenotipagem forense por DNA.

Justificativa

2. Justificativa

A determinação da relação entre o gene *SLC24A5*, seus SNPs e seu produto proteico, NCKX5, com fenótipos de pigmentação de uma população é conhecimento de grande importância para a área forense, que busca fazer uso deste e de outros genes relacionados a características externas visíveis no processo de predição do perfil morfológico de um suspeito cujo material genético tenha sido encontrado em uma cena de crime (Caliebe, et al. 2017).

Os sistemas IrisPlex e HIrisPlex, criados com este objetivo, não abrangem a grande variedade de fenótipos intermediários encontrados em populações miscigenadas, como cabelos castanhos, olhos castanhos e verdes e pele tipo III e IV (escala de Fitzpatrick) (Walsh, et al. 2013). Como já comprovado por nosso grupo de pesquisa, estes sistemas são de pouca utilidade no nosso país (Carratto 2017), que possui um grande histórico de migrações acarretando em uma população miscigenada e altamente variada em fenótipos de pigmentação. Outros polimorfismos genéticos envolvidos no processo de produção de melanina ainda não foram descritos.

Nenhum estudo publicado até o presente momento se aproximou do nível de abrangência e detalhamento empregados neste trabalho para se avaliar a diversidade do gene *SLC24A5*. A identificação de SNPs associados à pigmentação também é de grande interesse para a medicina. Muitas doenças que ocasionam hiper ou hipopigmentação como vitiligo, melasma e albinismo oculocutâneo, estão relacionadas a mutações em genes que desempenham algum papel na biossíntese de melanina. O gene *SLC24A5*, por exemplo, é conhecido por sua associação com um tipo de albinismo oculocutâneo, OCA6 (Bertolotti, et al. 2016; Morice-Picard, et al. 2014). A elucidação dos mecanismos causadores dessas doenças também pode auxiliar e direcionar tratamentos, ampliando a qualidade de vida destes pacientes.

Hipótese

3. Hipótese

Sabe-se que o gene *SLC24A5* está diretamente envolvido no processo de biossíntese de melanina. Este estudo tem por hipótese que SNPs presentes nas regiões regulatórias e codificantes deste gene influenciam no referido processo e, consequentemente, estão associados a diferentes fenótipos de pigmentação humana.

Objetivos

4. Objetivos

4.1. Objetivos gerais

Avaliar a diversidade genética das regiões regulatórias e codificantes do gene *SLC24A5* e identificar SNPs que estejam relacionados à pigmentação da pele, olhos e cabelos em amostra populacional da região de Ribeirão Preto.

4.2. Objetivos específicos

- Estimar frequências de alelos, haplótipos e genótipos relacionados aos SNPs das regiões promotora, codificante e 3'UTR do gene *SLC24A5*;
- Relacionar alelos, genótipos e haplótipos aos fenótipos apresentados pelos indivíduos amostrados;
- Avaliar o Desequilíbrio de Ligação entre sítios de variação nas regiões regulatórias e exônicas deste gene;
- Identificar computacionalmente miRNAs que apresentem função de reguladores pós-transcricionais deste gene, avaliando a interação de tais miRNAs com a presença de variantes na região 3'UTR;
- Identificar Fatores de Transcrição que atuem com reguladores de expressão e suas respostas a variantes nas regiões regulatórias.

Material e Métodos

5. Material e Métodos

5.1. Obtenção da amostra populacional de Ribeirão Preto (Anexo A)

Conforme proposto no projeto de pesquisa aprovado pelo Comitê de Ética em Pesquisa da FFCLRP/USP (Processo CEP – FFCLRP – CAAE n.25696413.7.0000.5407), os doadores assinaram o Termo de Consentimento Livre e Esclarecido (Anexo A) e receberam uma cópia assinada pelos pesquisadores responsáveis.

Os 407 participantes deste estudo foram recrutados principalmente no Hemocentro de Ribeirão Preto – SP, e em outras dependências do *campus* da USP de Ribeirão Preto. Cada doador forneceu 10 mL de sangue, coletado 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. O sangue foi armazenado em tubos *Vacutainer*® com anticoagulante EDTA a 4°C até a etapa de extração.

Além do sangue coletado, cada doador respondeu a um questionário, onde foram obtidas informações sobre suas características de pigmentação (autoclassificação), as quais também foram avaliadas por dois integrantes do Laboratório de Pesquisas Forenses e Genômicas – LPFG (heteroclassificação). Foram avaliadas a cor dos olhos (azul, verde mel, castanho-claro, castanho-escuro ou preto), homogeneidade da cor dos olhos nos primeiros 20 anos de vida e heterocromia, a cor natural dos cabelos (ruivo, loiro-claro, loiro-escuro, castanho-claro, castanho-escuro ou preto), homogeneidade da cor nos primeiros 20 anos de vida e porcentagem de cabelos grisalhos, a presença de pintas, manchas e sardas e a cor da pele segundo a escala de Fitzpatrick (Fitzpatrick 1988), que classifica os tons de pele em seis tipos considerando também sua capacidade de bronzeamento.

Os doadores também forneceram informações sobre a ancestralidade de seus avós (europeia, ameríndia, africana, asiática ou brasileira) e a sua ancestralidade e de seus pais segundo o IBGE (brancos, pretos, pardos, amarelos ou indígenas).

Foram obtidas imagens fotográficas dos olhos, cabelos e regiões da pele muito e pouco expostas ao sol, isso é, região frontal do rosto e região interna dos braços, respectivamente, utilizando a câmera SX50HS (Canon U.S.A., Inc.), com as seguintes configurações: abertura, $f = 3,4$, *shutterspeed* = 1/10 e ISO = 100. A distância média entre a câmera e os doadores foi aproximadamente sete centímetros.

Por fim, a fim de mensurar o índice de melanina e hemoglobina, foram obtidos espectros das mesmas regiões da pele, muito e pouco expostas ao sol, utilizando o espectrofotômetro CM-2500d e o software *SkinAnalysis* (Konica Minolta Holdings, Inc.).

5.2. Obtenção de outras amostras populacionais (Projeto 1000 Genomes)

Os dados do Projeto 1000 Genomes foram obtidos acessando o 1000 Genomes Browser (Fase 3) (<https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>). Dessa forma, foi obtido um arquivo *VCF (.vcf) correspondente à região avaliada (chr15: 48.411.008 – 48.437.098) por este estudo contendo os genótipos de 2504 indivíduos de 26 populações distribuídas em cinco grupos populacionais, conforme demonstra a Tabela 1.

Tabela 1 – Populações avaliadas pelo Projeto *1000 Genomes* (Fase 3) distribuídas conforme os cinco grupos populacionais amostrados. **Fonte:** (<https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>).

<i>Grupo populacional</i>	<i>Sigla (n)</i>	<i>População</i>	<i>Sigla (n)</i>
Africano	AFR (504)	Esan na Nigéria	ESN (99)
		Gambianos nas divisões ocidentais de Gâmbia	GWD (113)
		Luhya em Webuye, Quênia	LWK (99)
		Mende em Serra Leoa	MSL (85)
		Yoruba em Ibadan, Nigéria	YRI (108)
Americano	AMR (504)	Afro-caribenhos em Barbados	ACB (96)
		Americanos com ancestralidade africana no sudoeste dos EUA	ASW (61)
		Colombianos de Medellín, Colômbia	CLM (94)
		Ancestralidade Mexicana em Los Angeles, EUA	MXL (64)
		Peruanos de Lima, Peru	PEL (85)
		Porto riquenhos de Porto Rico	PUR (104)
Leste asiático	EAS (504)	Chineses Dai em Xishuangbanna, China	CDX (93)
		Chineses Han em Beijing, China	CHB (103)
		Chineses Han do sul	CHS (105)
		Japoneses em Tóquio, Japão	JPT (104)
		Kinh em Cidade de Ho Chi Minh, Vietnã	KHV (99)
Europeia	EUR (503)	Residentes de Utah (CEPH) com ancestralidade da Europa do Norte e Ocidental	CEU (99)
		Finlandeses na Finlândia	FIN (99)
		Britânicos na Inglaterra e Escócia	GBR (91)
		População ibérica na Espanha	IBS (107)
		Toscans na Itália	TSI (107)
		Bengali em Bangladesh	BEB (86)
Sul asiática	SAS (489)	Índios Gujarati de Houston, Texas	GIH (103)
		Índios Telugu do Reino Unido	ITU (102)
		Punjabi de Lahore, Paquistão	PJL (96)
		Tamil do Sri Lanka, do Reino Unido	STU (102)

5.3.Procedimentos experimentais

5.3.1. Extração (Anexo B – item 1)

A extração de DNA foi realizada a partir do sangue total, utilizando o método de *salting out* (Miller, et al. 1988) com modificações de acordo com o protocolo descrito no Anexo B – item 1. O DNA obtido foi ressuspensão em 200 µL de água e armazenado a -20°C.

5.3.2. Quantificação e verificação de qualidade do DNA genômico (Anexo B – item 2)

O DNA foi diluído dez vezes em água e posteriormente quantificado utilizando três metodologias. Foi realizada eletroforese em gel de agarose a fim de determinar eventuais degradações das amostras, espectrofotometria (*NanoDrop 2000 UV-Vis Spectrophotometer, Thermo Fisher Scientific Inc.*) com o intuito de determinar a presença de proteínas contaminantes desta amostra, e fluorimetria (*Qubit® 2.0 Fluorometer, Thermo Fisher Scientific Inc.*), a fim de determinar com acurácia a concentração do DNA genômico e preparar alíquotas de concentração 5ng/µL utilizadas no preparo de bibliotecas .

5.3.3. Preparo e validação de bibliotecas (Anexo B – itens 3 a 6)

O preparo de bibliotecas foi realizado com o uso do kit *HaloPlex Target Enrichment System (Agilent Technologies, Inc.)* customizado, ou seja, com sondas desenhadas a partir de um ensaio *in silico (SureDesign, Agilent Technologies, Inc.)* a fim de capturar e amplificar regiões contidas em dez genes relacionados à melanogênese. Para o gene *SLC24A5*, aqui estudado, as sondas permitem a cobertura das regiões regulatórias e exônicas (bem como porções intrônicas flanqueadoras) observadas na Figura 4.

Figura 4 – Captura de tela obtida no *Integrative Genomics Viewer 2.3.75* demonstrando as porções do gene *SLC24A5* cobertas pelo kit *HaloPlex Target Enrichment System* (Agilent Technologies, Inc.). A figura é dividida em dois quadrantes: o primeiro apresenta o cromossomo 15, onde a barra vermelha indica a posição do gene *SLC24A5*; o segundo quadrante apresenta em verde as regiões cobertas, os blocos azuis que representam os éxons deste gene e as linhas azuis que representam seus íntrons.

As bibliotecas foram preparadas 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 B. O processo consiste em cinco passos:

- a) Digestão das amostras por pares específicos de enzimas de restrição;
- b) Hibridização dos fragmentos alvo com sondas biotiniladas que incluem índices específicos para cada indivíduo;
- c) Captura dos fragmentos hibridizados utilizando *beads* magnéticas recobertas por estreptavidina;
- d) Amplificação dos fragmentos alvo por PCR;
- e) Purificação das bibliotecas.

Em seguida as bibliotecas foram validadas e quantificadas utilizando o sistema de eletroforese capilar *2100 BioAnalyzer* (Agilente Technologies, Inc.) e fluorimetria (*Qubit® 2.0 Fluorometer*, Thermo Fisher Scientific Inc.), especificados no item 6 do Anexo B.

5.3.4. Sequenciamento (Anexo B – item 7)

Após as validações, foi preparado um *pool* contendo no máximo 96 amostras para sequenciamento simultâneo, de forma que todas as amostras possuíssem concentrações próximas, tendo assim representatividades parecidas. Ao final, o *pool* foi diluído de forma a obter 600 µL de uma solução com concentração final de 16 pM. Essa concentração deve ser observada com cautela, pois concentrações maiores podem gerar uma alta clusterização, o que impediria a interpretação do sinal, e concentrações menores não atingiriam a sensibilidade mínima do equipamento.

Em seguida, esta solução foi inserida no cartucho *MiSeq Reagent Kit V3 (600 cycle)* para sequenciamento no *MiSeq Personal Sequencer* (Illumina Inc.) no modo *paired-end*, procedimento realizado de acordo com as instruções do fabricante e descrito no item 7 do Anexo B.

5.4.Procedimentos computacionais (Anexo C)

O protocolo encontrado no Anexo C descreve detalhadamente os procedimentos utilizados para tratamento e análise de dados sequenciamento. A seguir cada passo realizado será brevemente exposto.

5.4.1. Trimagem e mapeamento

O sequenciamento é realizado de forma que a adição de bases nitrogenadas modificadas com fluoróforos gere um sinal captado pelo detector. Os sinais são armazenados computacionalmente e podem ser obtidos, acessando o próprio sequenciador ou a plataforma *Basespace* (<https://basespace.illumina.com>), na forma de arquivos FastQ. Estes arquivos contêm as *reads* sequenciadas e informações de probabilidade de erro na determinação de cada uma das bases nitrogenadas (*Q-score*).

Entre as bases nitrogenadas sequenciadas ainda estão os adaptadores, adicionados durante o preparo de bibliotecas, os quais devem ser removidos, ou seja, trimados. Para removê-los utilizamos a ferramenta cutadapt (Martin 2011).

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

5.4.2. Obtenção de genótipos

A ferramenta HaplotypeCaller do pacote GATK (DePristo, et al. 2011; McKenna, et al. 2010; Van der Auwera, et al. 2013) foi utilizada para encontrar pontos de variação nas regiões promotora, codificante e 3'UTR, gerando um arquivo GVCF (*.g.vcf) que contém os genótipos obtidos para cada ponto de variação em cada um dos 407 indivíduos amostrados.

O software VCFx (Castelli, et al. 2015; Lima, et al. 2016) foi utilizado para obter informações de qualidade sobre as *reads* nos pontos de variação encontrados. Este software dispõe de duas ferramentas para realizar esta análise, checkpl e checkad, que interrogam alguns alelos substituindo-os por pontos. As interrogações realizadas por cada ferramenta foram verificadas manualmente utilizando o software *Integrative Genomics Viewer 2.3.75* (Robinson, et al. 2011; Thorvaldsdottir, et al. 2013), que nos mostra cada *read* alinhada a uma determinada região para determinado indivíduo. Com isso, é possível avaliar e eventualmente efetuar a correção de cada ponto de variação interrogado, de acordo com regras estabelecidas em função da quantidade e qualidade de *reads* neste ponto, descritas no Anexo C.

5.4.3. Obtenção de haplótipos (Anexo D – item 2)

O par de haplótipos de cada indivíduo foi inferido computacionalmente de acordo com o item 2 do Anexo D, utilizando o software PHASE 2.1.1 (Stephens e Donnelly 2003; Stephens, et al. 2001), o qual estima a probabilidade das possíveis combinações de haplótipos de cada indivíduo utilizando método Bayesiano. Com base nos sítios de variação identificados, o software analisa os possíveis pares de haplótipos de cada indivíduo e calcula a probabilidade das inferências destes haplótipos estarem corretas.

Neste estudo foi empregada a abordagem aqui designada como Faseamento Fragmentado, onde as regiões promotora e CDS foram divididas e os haplótipos inferidos separadamente. Os haplótipos obtidos para a região promotora continham nove sítios de variação, enquanto os haplótipos obtidos para a região CDS continham 22 sítios de variação.

A metodologia desta abordagem está descrita em detalhes no Anexo D com algumas modificações. Utilizou-se como probabilidade mínima de inferência 95%, ou seja, aqueles haplótipos com probabilidade inferior a 95% foram considerados incertos. Além disso, a nomenclatura dos haplótipos foi estabelecida simplesmente de acordo com a frequência dos mesmos.

5.4.4. Análises estatísticas (Anexos D e E)

Utilizando indivíduos não aparentados com haplótipos definidos e o software Arlequin 3.5.2.2 (Excoffier e Lischer 2010), foram estimadas frequências alélicas, frequências haplotípicas, diversidade genética, heterozigose esperada e observada, probabilidade de aderência ao Equilíbrio de Hardy-Weinberg (p_{EHW}) e distâncias genéticas em relação às 26 populações estudadas pelo Projeto *1000 Genomes* (Genomes Project, et al. 2015) considerando apenas os indivíduos coletados aleatoriamente. Os valores de distâncias genéticas foram avaliados mais profundamente por gráficos de *Multidimensional Scaling* utilizando o software IBM SPSS Statistics 20 (IBM 2011). O Desequilíbrio de Ligação (DL) foi avaliado pelo software Haploview 4.2 (Barrett, et al. 2005) para os 31 pontos de variação encontrados em toda a extensão do gene *SLC24A5*.

As associações entre alelos, genótipos ou haplótipos e fenótipos de pigmentação não foram realizadas de acordo com o item 5 do Anexo D, mas sim com o auxílio do software R Studio 1.1.456 (RStudio 2015) e um script desenvolvido por nosso grupo de pesquisa, e avaliadas segundo o Teste Exato de Fisher bi-caudal e valores de *Odds Ratio* com Intervalo de Confiança (IC) de 95%. Foram consideradas significativas as associações com valores de $p < 0,05$ e fortes com $OR > 10$ e $IC\ 95\% > 1$. As associações encontradas foram avaliadas segundo o impacto na função da proteína ou na regulação da expressão gênica utilizando o software Haploreg v4.1 (Ward e Kellis 2012), para pontos de variação em regiões regulatórias, e o software SNPEff (Cingolani, et al. 2012) e o Ensembl Browser (Zerbino, et al. 2018) para pontos em regiões codificantes.

O software mirhunt (Porto, et al. 2015) foi utilizado para estudar a existência de sítios de ligação de miRNA na região 3'UTR do gene *SLC24A5* de acordo com o Anexo E. Este software utiliza as ferramentas RNAhybrid (Rehmsmeier, et al. 2004), miRanda (John, et al. 2004) e intaRNA (Busch, et al. 2008) e concede três valores para esta avaliação: especificidade (S), estabilidade ou energia de ligação (E) e uma combinação dos valores anteriores (ES). Como não foram encontrados sítios da variação na região 3'UTR para a amostra obtida em Ribeirão Preto e pelo Projeto *1000 Genomes*, foi avaliado apenas o haplótipo referência.

A variação do índice de melanina e a relação com os sítios de variação associados foi observada utilizando os testes não paramétricos de Kruskal-Wallis e Mann-Whitney, por meio do software IBM SPSS Statistics 20.

Resultados

6. Resultados

6.1. Amostra populacional

Foram amostrados 407 indivíduos com idades entre 18 e 80 anos, sendo 195 do sexo feminino e 212 do sexo masculino. Em relação à pigmentação da pele, 61 indivíduos apresentaram pele escura (V e VI na escala Fitzpatrick), 162 apresentaram pele intermediária (III e IV na escala Fitzpatrick) e 184 apresentaram pele clara (I e II na escala Fitzpatrick). Quanto à pigmentação dos olhos, 180 indivíduos possuem olhos castanho-escuros, 90 possuem olhos castanho-claros, 20 possuem olhos mel, 75 olhos verdes e 42 olhos azuis. Em relação aos cabelos, 99 indivíduos possuem cabelos pretos, 149 possuem cabelos castanho-escuros, 76 possuem cabelos castanho-claros, 41 loiro-escuros, 17 loiro-claros e 11 ruivos. Não foi possível determinar a cor dos cabelos de 14 indivíduos, devido à calvície ou tintura. Quanto à ancestralidade do IBGE, 275 indivíduos foram classificados como brancos, 73 indivíduos foram classificados como pardos, 46 como pretos e 13 como amarelos. As características desta amostra populacional estão sumarizadas nos gráficos apresentados na Figura 5.

Figura 5 – Características da amostra populacional obtida em Ribeirão Preto de acordo com os fenótipos de pigmentação da pele, dos olhos e dos cabelos e a ancestralidade dos indivíduos segundo o IBGE.

6.2. Sequenciamento

As 407 amostras obtidas foram avaliadas em sete sequenciamentos, sendo que algumas amostras foram sequenciadas mais de uma vez. A Tabela 2 apresenta o número de amostras por sequenciamento (n) e alguns parâmetros de qualidade obtidos em cada sequenciamento. O alto número de *reads* obtido em cada sequenciamento reflete o sucesso desta etapa. Os valores de %PF (porcentagem de *clusters* do sequenciamento que passam pelo filtro de qualidade) também demonstram o sucesso dos sequenciamentos, sendo sempre maior que 75% e média igual a 87,16%. A porcentagem média de bases sequenciadas com PHRED 30 (probabilidade de erro equivalente a 1 em 1000) ou superior (%Q30 Médio) teve valor próximo a 55%, com média igual a 57,22%. O aparente baixo valor deste parâmetro não reflete adequadamente a qualidade do sequenciamento, como explicado Anexo C.

Tabela 2 – Dados de número de *reads*, aproveitamento e qualidade obtidos nos sete ensaios de sequenciamento realizados e a média de cada parâmetro.

<i>Sequenciamento</i>	<i>Nº de reads (milhões)</i>	<i>% PF</i>	<i>% Q30 Médio</i>
Sequenciamento 1 (<i>n</i> = 48)	48,48	78,79	55,65
Sequenciamento 2 (<i>n</i> = 95)	47,67	82,08	57,80
Sequenciamento 3 (<i>n</i> = 87)	21,32	96,72	62,67
Sequenciamento 4 (<i>n</i> = 80)	45,12	84,07	54,96
Sequenciamento 5 (<i>n</i> = 60)	43,60	85,74	56,28
Sequenciamento 6 (<i>n</i> = 35)	55,50	90,55	55,37
Sequenciamento 7 (<i>n</i> = 95)	41,46	92,17	57,80
Média	43,31	87,16	57,22

6.3.Variant calling

Utilizando a ferramenta *HaplotypeCaller* (GATK) foram encontrados 65 sítios de variação em 407 indivíduos, sendo 20 na região promotora e 45 na região CDS; não foram encontrados sítios de variação na região 3'UTR. A qualidade dos genótipos determinados em cada sítio de variação foi verificada utilizando as ferramentas *checkpl* e *checkad* da rotina VCFx. Estas ferramentas interrogam alelos de cada indivíduo de acordo com regras estabelecidas no Anexo C (item 1.3.2), considerando a cobertura total e a proporção de *reads* apresentando alelos referência e alternativos em determinado ponto de variação. Os sítios interrogados foram verificados visualmente utilizando o software *Integrative Genomics Viewer 2.3.75*.

Sete indivíduos que tiveram quantidade de *double missing* (ou seja, ambos os alelos interrogados em um ponto de variação) maior ou igual a 10% foram excluídos. Dentre os 65 pontos de variação, 34 tratavam-se de *singletons*, ou seja, sítios que apresentavam o alelo alternativo em apenas uma das amostra analisadas, e foram excluídos: 11 deles na região promotora e 23 na região CDS. Ao final das correções de interrogações e remoção de *singletons* um arquivo *VCF (.vcf) final foi criado contendo 400 indivíduos e 31 sítios de variação revisados: nove na região promotora, 20 em regiões intrônicas (sendo dois ainda não descritos e a grande maioria no segundo e maior íntron deste gene, com aproximadamente 12 kb) e dois em regiões exônicas (Tabela 3).

Tabela 3 – Sítios de variação (excluindo-se *singletons*) encontrados na amostra populacional de Ribeirão Preto, suas posições no cromossomo 15 utilizando o genoma hg19 como referência e regiões em que se encontram no gene *SLC24A5*.

<i>Sítio de variação</i>	<i>Posição do cromossomo 15</i>	<i>Região no gene</i>
rs113364332	48.410.353	Promotora
rs10162789	48.410.384	Promotora
rs142234630	48.410.517	Promotora
rs79875456	48.411.605	Promotora
rs553733344	48.411.795	Promotora
rs55728404	48.411.805	Promotora
rs2675346	48.411.821	Promotora
rs78032672	48.412.360	Promotora
rs59973476	48.412.961	Promotora
rs150119218	48.414.553	Íntron 2
rs11344363	48.414.604	Íntron 2
rs550643847	48.414.901	Íntron 2
rs2433354	48.414.969	Íntron 2
rs2459391	48.415.068	Íntron 2
rs77575793	48.415.133	Íntron 2
rs570212959	48.415.157	Íntron 2
rs79856063	48.415.196	Íntron 2
rs2433355	48.415.261	Íntron 2
chr 15:48415361	48.415.361	Íntron 2
chr15:48415656	48.415.656	Íntron 2
rs79338096	48.415.754	Íntron 2
rs76084158	48.415.765	Íntron 2
rs189640907	48.415.770	Íntron 2
rs2433356	48.416.360	Íntron 2
rs138961421	48.416.756	Íntron 2
rs16960620	48.416.899	Íntron 2
rs1426654	48.426.484	Éxon 3
rs57342157	48.426.800	Íntron 4
rs538570737	48.427.312	Íntron 5
rs142049412	48.428.889	Éxon 6
rs76547866	48.433.306	Íntron 7

Os SNPs rs2555364, para o qual já foram descritas associações com fenótipos de pigmentação (Giardina, et al. 2008a), e rs17426596, para o qual foram encontradas

associações com fenótipos de pigmentação e melanoma, porém não significativas (Nan, et al. 2009), não foram cobertos pelo nosso ensaio, e por isso não foi possível encontrar suas variações e associações com fenótipos de pigmentação no presente estudo.

6.4. Frequências alélicas, heterozigose, frequências haplotípicas, e diversidade haplotípica

Durante a etapa de inferência de haplótipos sete indivíduos não atingiram o mínimo de probabilidade de inferência estabelecido em 95%, e foram removidos. Excluindo seis indivíduos aparentados presentes em nossa amostra e considerando os 387 indivíduos não aparentados e com haplótipos definidos, calculou-se a frequência alélica de cada um dos 31 sítios de variação identificados (nove na região promotora e 22 na região CDS), apresentadas na Tabela 4. Dentre estes sítios, apenas nove são polimórficos, ou seja, a frequência do alelo menos frequente é maior que 1%.

Foram calculados também os valores de heterozigose observada (H_o) e esperada (H_e), além da probabilidade de aderência ao Equilíbrio de Hardy-Weinberg (p_{EHW}) considerando indivíduos coletados aleatoriamente, também apresentadas na Tabela 4. Dentre os nove sítios polimórficos encontrados, quatro deles desviam significativamente do Equilíbrio de Hardy-Weinberg ($p < 0,05$): rs2675346, rs2433354, rs2433356, rs1426654. Para os quatro SNPs citados a heterozigose esperada é maior do que a heterozigose observada, o que aponta uma deficiência de heterozigotos.

Tabela 4 – Frequências de alelos alternativos e referências dos 31 sítios de variação avaliados, seus valores de heterozigose observada e esperada e aderência ao Equilíbrio de Hardy-Weinberg para a população de Ribeirão Preto. Em negrito estão os sítios de variação polimórficos. Sublinhados estão os sítios de variação que desviam do Equilíbrio de Hardy-Weinberg.

<i>Sítio de variação</i>	<i>Frequência</i>		<i>H_o</i>	<i>H_e</i>	<i>pEHW</i>
	<i>Alelo referência</i>	<i>Alelo alternativo</i>			
rs113364332	0,9975	0,0025	0,0051	0,0051	1,0000
rs10162789	0,9962	0,0038	0,0076	0,0076	1,0000
rs142234630	0,9962	0,0038	0,0076	0,0076	1,0000
rs79875456	0,9847	0,0153	0,0305	0,0301	1,0000
rs553733344	0,9517	0,0483	0,0967	0,0921	1,0000
rs55728404	0,9529	0,0471	0,0942	0,0898	1,0000
<u>rs2675346</u>	<u>0,8588</u>	<u>0,1412</u>	<u>0,1908</u>	<u>0,2429</u>	<u>0,0001</u>
rs78032672	0,9975	0,0025	0,0051	0,0051	1,0000
rs59973476	0,9975	0,0025	0,0051	0,0051	1,0000
rs150119218	0,9962	0,0038	0,0076	0,0076	1,0000
rs11344363	0,9936	0,0064	0,0127	0,0127	1,0000
rs550643847	0,9975	0,0025	0,0051	0,0051	1,0000
<u>rs2433354</u>	<u>0,8588</u>	<u>0,1412</u>	<u>0,1908</u>	<u>0,2429</u>	<u>0,0001</u>
rs2459391	0,9517	0,0483	0,0967	0,0921	1,0000
rs77575793	0,9924	0,0076	0,0153	0,0152	1,0000
rs570212959	0,9962	0,0038	0,0076	0,0076	1,0000
rs79856063	0,9962	0,0038	0,0076	0,0076	1,0000
rs2433355	0,0089	0,9911	0,0178	0,0177	1,0000
chr15:48415361	0,9924	0,0076	0,0153	0,0152	1,0000
chr15:48415656	0,9975	0,0025	0,0051	0,0051	1,0000
rs79338096	0,9720	0,028	0,0560	0,0545	1,0000
rs76084158	0,9975	0,0025	0,0051	0,0051	1,0000
rs189640907	0,9949	0,0051	0,0102	0,0101	1,0000
<u>rs2433356</u>	<u>0,9211</u>	<u>0,0789</u>	<u>0,1221</u>	<u>0,1455</u>	<u>0,0058</u>
rs138961421	0,9962	0,0038	0,0076	0,0076	1,0000
rs16960620	0,9962	0,0038	0,0076	0,0076	1,0000
<u>rs1426654</u>	<u>0,7443</u>	<u>0,2557</u>	<u>0,2570</u>	<u>0,3811</u>	<u>0,0000</u>
rs57342157	0,9975	0,0025	0,0051	0,0051	1,0000
rs538570737	0,9975	0,0025	0,0051	0,0051	1,0000
rs142049412	0,9962	0,0038	0,0076	0,0076	1,0000
rs76547866	0,9924	0,0076	0,0153	0,0152	1,0000

Foram encontrados 58 haplótipos (Tabela 5) distribuídos entre a população de Ribeirão Preto (RBP) e os cinco grupos populacionais amostrados pelo Projeto *1000 Genomes*: africano (AFR), americano (AMR), leste asiático (EAS), europeu (EUR) e sul asiático (SAS). As frequências haplotípicas observadas nas populações estão apresentadas na Tabela 6.

Dentre os 58 haplótipos, 39 estão presentes em Ribeirão Preto e 18 destes não foram encontrados em nenhum outro grupo populacional. O haplótipo 1 (HAP 01) apresentou alta frequência na população de Ribeirão Preto (70,10%) e nos grupos populacionais sul asiático (SAS) (67,83%) e, especialmente, europeu (EUR) (96,22%). O grupo populacional europeu (EUR) apresentou baixa variabilidade com apenas 14 haplótipos, enquanto o grupo populacional africano (AFR) apresentou alta variabilidade com 24 haplótipos, sendo a maior frequência igual a 25,10% para HAP 02.

Tabela 5 – Haplótipos do gene *SLC24A5* encontrados na região de Ribeirão Preto indicando o alelo que possui em cada sítio de variação identificado. Os pontos observados nas diferentes posições indicam que o haplótipo em questão possui a base presentes na sequência de referência.

ID	rs113364332	rs10162789	rs142234630	rs79875456	rs553733344	rs55728404	rs2675346	rs78032672	rs59973476	rs150119218	rs11344363	rs550643847	rs2433354	rs2459391	rs77575793	rs570212959	rs79856063	rs2433355	chr15: 48415361	chr15: 48415656	rs79338096	rs76084158	rs189640907	rs2433356	rs138961421	rs16960620	rs1426654	rs57342157	rs538570737	rs142049412	rs76547866
REF	T	T	G	C	C	T	C	A	A	G	CT	G	C	A	C	A	T	C	AAT	A	A	T	A	G	A	A	A	G	G	G	A
HAP 01	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 02	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	.	.
HAP 03	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 04	.	.	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	C	.	.	A	.	.	G	.	.	.	.
HAP 05	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 06	.	.	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 07	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 08	.	.	.	G	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	.	.
HAP 09	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	A	.	.	.	.	.	.	.	.	.	.	.	.
HAP 11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	T
HAP 12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.
HAP 13	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 14	.	.	.	G	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 15	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	.	.
HAP 16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	A	.	.	.
HAP 17	.	.	.	.	.	.	T	.	.	.	C	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	G	.	.	.	.
HAP 18	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	AC	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.
HAP 20	.	.	.	.	.	.	T	.	.	.	C	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	.	.
HAP 21	C	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	.	.

ID	rs113364332	rs10162789	rs142234630	rs79875456	rs553733344	rs55728404	rs2675346	rs78032672	rs59973476	rs150119218	rs11344363	rs550643847	rs2433354	rs2459391	rs77575793	rs570212959	rs79856063	rs2433355	chr15: 48415361	chr15: 48415656	rs79338096	rs76084158	rs189640907	rs2433356	rs138961421	rs16960620	rs1426654	rs57342157	rs538570737	rs142049412	rs76547866
REF	T	T	G	C	C	T	C	A	A	G	CT	G	C	A	C	A	T	C	AAT	A	A	T	A	G	A	A	A	G	G	G	A
HAP 22	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 23	.	.	.	.	.	.	T	C	G	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 24	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	C	.	.	.	G	.	.	.	.	.
HAP 25	.	.	.	G	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	A	.	.	.	.	.	.	G	.	.	.	A	.
HAP 27	.	.	.	.	.	.	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 28	.	.	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	T	A	.	.	G	.	.	.	.
HAP 29	.	G	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	C	T	.	.	.	.	.	A	C	.	G	.	.	A	.
HAP 30	.	.	.	.	CA	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 31	.	G	.	G	CA	G	T	.	.	.	.	.	T	G	.	.	C	T	.	.	.	.	.	A	C	.	G	.	.	.	.
HAP 32	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	G	.	.	.	.	.	G	.	.	.	.	T
HAP 33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	G	.	.	.	.	.	G	.	.	.	.	.
HAP 34	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	T
HAP 35	.	.	.	.	.	.	T	.	.	A	.	.	T	.	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 36	.	G	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	C	T	.	.	.	.	.	A	C	.	G	.	.	.	.
HAP 37	.	.	.	.	.	.	T	.	.	.	C	.	T	.	.	.	.	T	.	.	.	C	.	.	.	G	G	.	.	.	.
HAP 38	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	AC	.	T	.	.	.	.	.	.	.	G	.	.	A	.	.
HAP 39	.	.	.	G	.	.	T	.	.	.	.	.	T	.	.	.	.	T	A	.	.	.	.	.	.	G	.	.	.	.	.
HAP 40	.	.	.	G	.	.	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 41	.	.	.	.	.	.	T	.	G	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 42	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	AC	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 43	.	.	.	.	.	.	T	.	.	.	.	.	T	G	.	.	.	T	.	.	C	.	.	A	.	.	G	.	.	.	.
HAP 44	.	.	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.

ID	rs113364332	rs10162789	rs142234630	rs79875456	rs553733344	rs55728404	rs2675346	rs78032672	rs59973476	rs150119218	rs11344363	rs550643847	rs2433354	rs2459391	rs77575793	rs570212959	rs79856063	rs2433355	chr15: 48415361	chr15: 48415656	rs79338096	rs76084158	rs189640907	rs2433356	rs138961421	rs16960620	rs1426654	rs57342157	rs538570737	rs142049412	rs76547866
REF	T	T	G	C	C	T	C	A	A	G	CT	G	C	A	C	A	T	C	AAT	A	A	T	A	G	A	A	A	G	G	G	A
HAP 45	.	.	.	.	.	G	T	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	.	.
HAP 46	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	A	.	.
HAP 47	.	.	.	.	.	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	.	.	A	.	.	G	.	.	.	.	.
HAP 48	.	.	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	C	.	.	.	.	G	.	C	.	.	.
HAP 49	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 50	.	.	.	.	CA	G	T	.	.	.	.	.	T	G	.	.	.	T	.	.	C	.	.	A	.	.	G	A	.	.	.
HAP 51	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	T	.	.	C	.	.	A	.	.	G	.	.	.	.
HAP 52	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	C	.	.	.	.	G	.	.	.	.
HAP 53	.	.	.	.	.	.	T	.	G	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	G	.	.	.	.
HAP 54	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	G	.	.	.	.
HAP 55	.	.	.	.	CA	G	T	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	.	G	.	.	.	.	.
HAP 56	.	.	.	.	CA	G	T	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	G	.	.	.	.
HAP 57	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.
HAP 58	.	.	.	.	.	.	T	.	.	.	.	.	T	.	.	.	.	T	.	.	C	.	.	A	.	.	G	.	.	.	.

Tabela 6 – Frequências dos haplótipos encontrados na região de Ribeirão Preto e nos grupos populacionais avaliados pelo Projeto *1000 Genomes*.

HAP	RBP (2n = 774)	AFR (2n = 1008)	AMR (2n = 1008)	EAS (2n = 1008)	EUR (2n = 1006)	SAS (2n = 978)
HAP 01	0,7010	0,0546	0,4395	0,0119	0,9622	0,6783
HAP 02	0,1069	0,2510	0,2321	0,4712	0,0020	0,1004
HAP 03	0,0356	0,1081	0,0675	0,1680		0,1004
HAP 04	0,0280	0,0833	0,0962	0,0606	0,0010	0,0113
HAP 05	0,0267	0,2024	0,0615			
HAP 06	0,0127	0,1250	0,0427	0,0020		0,0533
HAP 07	0,0089				0,0030	
HAP 08	0,0089	0,0734	0,0179			
HAP 09	0,0076		0,0050		0,0129	0,0031
HAP 10	0,0051					
HAP 11	0,0051		0,0010		0,0119	0,0020
HAP 12	0,0038		0,0010		0,0030	0,0010
HAP 13	0,0038				0,0040	
HAP 14	0,0025					
HAP 15	0,0025	0,0020	0,0010			
HAP 16	0,0025					
HAP 17	0,0025			0,1252		
HAP 18	0,0025	0,0089	0,0040			
HAP 19	0,0025					
HAP 20	0,0025	0,0129	0,0089	0,0934		0,0400
HAP 21	0,0025	0,0030				
HAP 22	0,0025		0,0020			
HAP 23	0,0025			0,0229		
HAP 24	0,0013					
HAP 25	0,0013					
HAP 26	0,0013					
HAP 27	0,0013	0,0040	0,0010			
HAP 28	0,0013					
HAP 29	0,0013					
HAP 30	0,0013					
HAP 31	0,0013					
HAP 32	0,0013					
HAP 33	0,0013					
HAP 34	0,0013					

HAP	RBP (2n = 774)	AFR (2n = 1008)	AMR (2n = 1008)	EAS (2n = 1008)	EUR (2n = 1006)	SAS (2n = 978)
HAP 35	0,0013					
HAP 36	0,0013	0,0397	0,0089			
HAP 37	0,0013					
HAP 38	0,0013					
HAP 39	0,0013					
HAP 40		0,0208	0,0040			
HAP 41		0,0020	0,0010	0,0060		0,0061
HAP 42		0,0010				
HAP 43		0,0010				
HAP 44		0,0010				
HAP 45		0,0010				
HAP 46		0,0010				
HAP 47		0,0010				
HAP 48		0,0010	0,0020			
HAP 49		0,0010		0,0020		
HAP 50		0,0010	0,0020			
HAP 51			0,0010			
HAP 52				0,0249		
HAP 53				0,0089		
HAP 54				0,0020		
HAP 55				0,0010		
HAP 56						0,0020
HAP 57						0,0010
HAP 58						0,0010

Foram calculadas as diversidades haplotípicas para a população de Ribeirão Preto e para os grupos populacionais do Projeto *1000 Genomes* (Tabela 7). Observa-se que a população de Ribeirão Preto apresenta baixa diversidade haplotípica, apesar da grande quantidade de haplótipos presentes. Este valor é decorrente da alta frequência de HAP 01, a maior frequência entre todos os haplótipos desta população. A alta frequência observada para o mesmo haplótipo nos grupos sul asiático e, em especial, europeu, também causaram uma diminuição expressiva em suas diversidades haplotípicas.

Tabela 7 – Diversidade haplotípica do gene *SLS24A5* calculada para a amostra populacional da região de Ribeirão Preto e para os grupos populacionais estudados pelo Projeto *1000 Genomes*.

<i>Grupo populacional</i>	<i>Diversidade haplotípica</i>
RBP	0,4945
AFR	0,8520
AMR	0,7337
EAS	0,7210
EUR	0,0738
SAS	0,5157

6.5.Diferenciação populacional

Os testes de diferenciação populacional aqui utilizados, índice F_{ST} e teste exato de diferenciação populacional baseado em frequências haplotípicas, são complementares, servindo para apoiar seus respectivos resultados. Independente do valor de F_{ST} , este deve ser interpretado considerando-se a significância estatística, isto é, o valor de probabilidade de não diferenciação. Considera-se que pequenas distâncias entre populações resultem em valores de F_{ST} significativos de até 0,05, distância intermediária entre 0,05 e 0,15, grande distância entre 0,15 e 0,25 e distância muito grande acima de 0,25 (Hartl 2010).

Para o teste exato de diferenciação são apresentados valores de probabilidade de não diferenciação para cada par de populações. Valores de probabilidade menores que 0,05 indicam diferenciação estatisticamente significativa.

As Tabelas 8, 9 e 10 apresentam, respectivamente, os valores de F_{ST} entre a amostra de Ribeirão Preto e cada população amostrada pelo Projeto *1000 Genomes*, os valores do teste exato de diferenciação entre as mesmas populações e os valores de F_{ST} entre a amostra de Ribeirão Preto e os grupos populacionais amostrados pelo projeto. O teste exato de diferenciação populacional baseado em frequências haplotípicas também foi realizado entre a população de Ribeirão Preto e os grupos populacionais do Projeto *1000 Genomes* e revelou que todos os pares de populações e grupos populacionais estudados diferem significativamente entre si ($p = 0,0000 \pm 0,0000$).

Tabela 8 – Índices de F_{ST} baseado em frequências haplotípicas estimados entre a amostra populacional de Ribeirão Preto e as populações amostradas pelo Projeto *1000 Genomes*.

		RBP	AFR					AMR					EAS					EUR					SAS					
			ESN	GWD	LWK	MSL	YRI	ACB	ASW	CLM	MXL	PEL	PUR	CDX	CHB	CHS	JPT	KHV	CEU	FIN	GBR	IBS	TSI	BEB	GIH	ITU	PJL	STU
RBP		-	0,44	0,42	0,35	0,44	0,45	0,39	0,29	0,01	0,05	0,15	0,00	0,35	0,34	0,31	0,32	0,33	0,08	0,08	0,08	0,09	0,08	0,06	0,06	0,01	0,00	0,09
AFR	ESN	0,44	-	0,01	0,02	0,00	0,00	0,00	0,04	0,31	0,25	0,16	0,40	0,15	0,14	0,16	0,18	0,15	0,59	0,59	0,58	0,60	0,59	0,24	0,56	0,32	0,40	0,19
	GWD	0,42	0,01	-	0,02	0,01	0,01	0,01	0,03	0,31	0,24	0,16	0,39	0,12	0,11	0,13	0,15	0,13	0,58	0,58	0,57	0,59	0,59	0,22	0,56	0,30	0,39	0,17
	LWK	0,35	0,02	0,02	-	0,02	0,02	0,00	0,00	0,24	0,17	0,08	0,32	0,08	0,08	0,08	0,11	0,08	0,53	0,53	0,52	0,54	0,53	0,16	0,50	0,23	0,32	0,11
	MSL	0,44	0,00	0,01	0,02	-	0,00	0,00	0,04	0,32	0,25	0,17	0,40	0,15	0,14	0,17	0,19	0,16	0,61	0,61	0,60	0,62	0,61	0,24	0,58	0,32	0,41	0,19
	YRI	0,45	0,00	0,01	0,02	0,00	-	0,00	0,05	0,32	0,25	0,16	0,40	0,14	0,13	0,15	0,18	0,15	0,59	0,59	0,58	0,60	0,59	0,24	0,56	0,32	0,41	0,19
AMR	ACB	0,39	0,00	0,01	0,00	0,00	0,00	-	0,02	0,26	0,20	0,12	0,35	0,12	0,11	0,12	0,15	0,12	0,55	0,55	0,54	0,56	0,55	0,19	0,52	0,26	0,35	0,14
	ASW	0,29	0,04	0,03	0,00	0,04	0,05	0,02	-	0,18	0,11	0,05	0,27	0,08	0,07	0,07	0,09	0,07	0,54	0,53	0,52	0,55	0,54	0,10	0,49	0,18	0,27	0,06
	CLM	0,01	0,31	0,31	0,24	0,32	0,32	0,26	0,18	-	0,01	0,08	0,01	0,27	0,26	0,24	0,26	0,25	0,16	0,16	0,16	0,17	0,16	0,03	0,12	0,01	0,02	0,05
	MXL	0,05	0,25	0,24	0,17	0,25	0,25	0,20	0,11	0,01	-	0,02	0,05	0,20	0,19	0,16	0,18	0,17	0,29	0,29	0,28	0,30	0,29	0,02	0,23	0,02	0,06	0,02
	PEL	0,15	0,16	0,16	0,08	0,17	0,16	0,12	0,05	0,08	0,02	-	0,14	0,11	0,10	0,07	0,10	0,09	0,39	0,38	0,37	0,40	0,39	0,04	0,34	0,08	0,16	0,02
	PUR	0,00	0,40	0,39	0,32	0,40	0,40	0,35	0,27	0,01	0,05	0,14	-	0,34	0,33	0,30	0,33	0,32	0,12	0,11	0,11	0,12	0,11	0,07	0,07	0,02	0,00	0,09
EAS	CDX	0,35	0,15	0,12	0,08	0,15	0,14	0,12	0,08	0,27	0,20	0,11	0,34	-	0,00	0,00	0,01	0,00	0,58	0,58	0,57	0,59	0,58	0,15	0,54	0,24	0,34	0,13
	CHB	0,34	0,14	0,11	0,08	0,14	0,13	0,11	0,07	0,26	0,19	0,10	0,33	0,00	-	0,01	0,01	0,00	0,55	0,55	0,54	0,56	0,55	0,14	0,51	0,23	0,32	0,13
	CHS	0,31	0,16	0,13	0,08	0,17	0,15	0,12	0,07	0,24	0,16	0,07	0,30	0,00	0,01	-	0,00	0,00	0,54	0,54	0,53	0,55	0,54	0,12	0,50	0,20	0,30	0,11
	JPT	0,32	0,18	0,15	0,11	0,19	0,18	0,15	0,09	0,26	0,18	0,10	0,33	0,01	0,01	0,00	-	0,00	0,56	0,56	0,55	0,57	0,57	0,14	0,52	0,22	0,32	0,12
	KHV	0,33	0,15	0,13	0,08	0,16	0,15	0,12	0,07	0,25	0,17	0,09	0,32	0,00	0,00	0,00	0,00	-	0,56	0,55	0,54	0,57	0,56	0,14	0,52	0,22	0,32	0,12
EUR	CEU	0,08	0,59	0,58	0,53	0,61	0,59	0,55	0,54	0,16	0,29	0,39	0,12	0,58	0,55	0,54	0,56	0,56	-	0,01	0,00	0,00	0,00	0,30	0,02	0,20	0,14	0,31
	FIN	0,08	0,59	0,58	0,53	0,61	0,59	0,55	0,53	0,16	0,29	0,38	0,11	0,58	0,55	0,54	0,56	0,55	0,01	-	0,01	0,01	0,00	0,29	0,01	0,19	0,14	0,30
	GBR	0,08	0,58	0,57	0,52	0,60	0,58	0,54	0,52	0,16	0,28	0,37	0,11	0,57	0,54	0,53	0,55	0,54	0,00	0,01	-	0,01	0,01	0,28	0,02	0,19	0,13	0,30
	IBS	0,09	0,60	0,59	0,54	0,62	0,60	0,56	0,55	0,17	0,30	0,40	0,12	0,59	0,56	0,55	0,57	0,57	0,00	0,01	0,01	-	0,00	0,30	0,02	0,20	0,15	0,32
	TSI	0,08	0,59	0,59	0,53	0,61	0,59	0,55	0,54	0,16	0,29	0,39	0,11	0,58	0,55	0,54	0,57	0,56	0,00	0,00	0,01	0,00	-	0,29	0,01	0,19	0,14	0,31
SAS	BEB	0,06	0,24	0,22	0,16	0,24	0,24	0,19	0,10	0,03	0,02	0,04	0,07	0,15	0,14	0,12	0,14	0,14	0,30	0,29	0,28	0,30	0,29	-	0,24	0,01	0,06	0,00
	GIH	0,06	0,56	0,56	0,50	0,58	0,56	0,52	0,49	0,12	0,23	0,34	0,07	0,54	0,51	0,50	0,52	0,52	0,02	0,01	0,02	0,02	0,01	0,24	-	0,15	0,09	0,26
	ITU	0,01	0,32	0,30	0,23	0,32	0,32	0,26	0,18	0,01	0,02	0,08	0,02	0,24	0,23	0,20	0,22	0,22	0,20	0,19	0,19	0,20	0,19	0,01	0,15	-	0,01	0,03
	PJL	0,00	0,40	0,39	0,32	0,41	0,41	0,35	0,27	0,02	0,06	0,16	0,00	0,34	0,32	0,30	0,32	0,32	0,14	0,14	0,13	0,15	0,14	0,06	0,09	0,01	-	0,09
	STU	0,09	0,19	0,17	0,11	0,19	0,19	0,14	0,06	0,05	0,02	0,02	0,09	0,13	0,13	0,11	0,12	0,12	0,31	0,30	0,30	0,32	0,31	0,00	0,26	0,03	0,09	-

Tabela 9 – Teste exato de diferenciação baseado em frequências haplotípicas estimado entre a amostra populacional de Ribeirão Preto e populações amostradas pelo Projeto *1000 Genomes*.

		RBP	AFR					AMR					EAS					EUR					SAS				
			ESN	GWD	LWK	MSL	YRI	ACB	ASW	CLM	MXL	PEL	PUR	CDX	CHB	CHS	JPT	KHV	CEU	FIN	GBR	IBS	TSI	BEB	GIH	ITU	PJL
RBP		-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,56	0,00	0,00	0,97	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
AFR	ESN	0,00	-	0,01	0,00	0,14	0,40	0,05	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	GWD	0,00	0,01	-	0,00	0,23	0,00	0,07	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	LWK	0,00	0,00	0,00	-	0,00	0,00	0,06	0,14	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	MSL	0,00	0,14	0,23	0,00	-	0,01	0,32	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	YRI	0,00	0,40	0,00	0,00	0,01	-	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
AMR	ACB	0,00	0,05	0,07	0,06	0,32	0,02	-	0,21	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	ASW	0,00	0,00	0,00	0,14	0,03	0,00	0,21	-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	CLM	0,56	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,00	0,00	0,27	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	MXL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	PEL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	PUR	0,97	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,27	0,00	0,00	0,00	-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EAS	CDX	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,48	0,12	0,19	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	CHB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,48	-	0,20	0,09	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	CHS	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,20	-	0,12	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	JPT	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,19	0,09	0,12	-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	KHV	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,86	0,39	0,50	0,20	-	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EUR	CEU	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,03	0,27	0,80	0,17	0,00	0,00	0,00	0,00	0,00
	FIN	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,03	-	0,02	0,03	0,18	0,00	0,12	0,00	0,00	0,00
	GBR	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,27	0,27	0,02	-	0,13	0,02	0,00	0,00	0,00	0,00	0,00
	IBS	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,80	0,80	0,03	0,13	-	0,58	0,00	0,00	0,00	0,00	0,00
	TSI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,17	0,18	0,02	0,58	-	0,00	0,02	0,00	0,00	0,00
SAS	BEB	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,00	0,10	0,00	0,03
	GIH	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,12	0,00	0,00	0,02	0,00	-	0,00	0,00	0,00
	ITU	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,10	0,00	-	0,00	0,01
	PJL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-	0,00
	STU	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,00	0,01	0,00	-

Tabela 10 – Índices de F_{ST} baseado em frequências haplotípicas estimados entre a amostra populacional de Ribeirão Preto e os grupos populacionais amostrados pelo Projeto *1000 Genomes*.

	RBP	AFR	AMR	EAS	EUR	SAS
RBP	-	0,3544	0,0830	0,2967	0,1369	0,0090
AFR	0,3544	-	0,1501	0,1336	0,5776	0,3138
AMR	0,0830	0,1501	-	0,1230	0,3013	0,0539
EAS	0,2967	0,1336	0,1230	-	0,5572	0,2442
EUR	0,1369	0,5776	0,3013	0,5572	-	0,1865
SAS	0,0090	0,3138	0,0539	0,2442	0,1865	-

O índice de F_{ST} mostrou que as populações MXL, CLM e PUR (grupo populacional americano) apresentam pequenas distâncias estimadas por F_{ST} em relação à população de Ribeirão Preto (0,0495, 0,0120 e -0,0004, respectivamente), assim como as populações PJI e ITU (grupo sul asiático), também com pequenas distâncias genéticas (0,0010 e 0,0129, respectivamente). Entretanto, embora pequenas, as distâncias entre RBP e MXL, CLM ou PJI são estatisticamente significantes. Todas as populações europeias apresentaram distâncias intermediárias e significantes ($p = 0,0000 \pm 0,0000$) em relação à população de Ribeirão Preto, enquanto as outras populações apresentaram distâncias grandes ou muito grandes e também estatisticamente significantes ($p = 0,0000 \pm 0,0000$). Observamos ainda que a população GIH apresenta pequenas distâncias em relação a todas as populações europeias, com índice de F_{ST} entre 0,0106 e 0,0206 ($p = 0,0000 \pm 0,0000$).

Já o teste exato de diferenciação demonstrou que a população de Ribeirão Preto não se diferencia significativamente apenas das populações CLM e PUR ($0,5607 \pm 0,0471$ e $0,9677 \pm 0,0068$).

Levando em consideração os grupos populacionais, o índice de F_{ST} mostra que os grupos africano e leste asiático apresentam distâncias muito grandes em relação à população de Ribeirão Preto, enquanto as populações americana e europeia apresentam distâncias intermediárias, como previsto pela composição da população brasileira. Entretanto, o grupo sul asiático demonstrou uma pequena distância em relação à população de Ribeirão Preto, o que não era esperado.

Construindo gráficos *Multidimensional Scaling* a partir dos valores F_{ST} obtidos é possível observar a distância de populações (Figuras 6) e grupos populacionais (Figura 7) mundiais (Projeto *1000 Genomes*) em relação à população de Ribeirão Preto em duas dimensões. Observa-se que as populações sul asiáticas se aproxima gradativamente das populações europeias, e, por consequência, se aproximam da população de Ribeirão Preto.

Figura 6 – Distribuição bidimensional das distâncias obtidas entre Ribeirão Preto e as populações avaliadas pelo Projeto *1000 Genomes*. A distribuição foi realizada de acordo com os valores de distância genética (F_{ST}) calculados.

Figura 7 – Distribuição bidimensional das distâncias obtidas entre Ribeirão Preto e os grupos populacionais avaliados pelo Projeto *1000 Genomes*. A distribuição foi realizada de acordo com os valores de distância genética (F_{ST}) calculados.

6.6. Desequilíbrio de Ligação

O software Haploview gerou a Figura 8, formada por quadrados que representam o desequilíbrio de ligação (D') entre os dois sítios de variação que o formam. Os blocos de haplótipos foram definidos de acordo com o método de Gabriel, et al. (2002). Os quadrados vermelhos indicam intenso DL, LOD ($Log\ of\ Odds$) ≥ 2 e $D' = 1$, os brancos indicam ausência de DL, $LOD < 2$ e $D' < 1$, e os azuis indicam fraco, porém completo DL, $LOD < 2$ e $D' = 1$.

Figura 8 – Representação do Desequilíbrio de Ligação (DL) observado no gene *SLC24A5* obtido de acordo com o método de Gabriel, et al. (2002). Os quadrados vermelhos indicam intenso DL, LOD ($Log\ of\ Odds$) ≥ 2 e $D' = 1$, os brancos indicam ausência de DL, $LOD < 2$ e $D' < 1$, e os azuis indicam fraco, porém completo DL, $LOD < 2$ e $D' = 1$. Encontramos um grande bloco de aproximadamente 14 kb em forte DL devido aos SNPs rs2675346 (7), rs2433354 (13), rs2433356 (24) e rs142665 (27). No canto inferior direito encontram-se as frequências dos haplótipos levando em conta os quatro sítios de variação.

Observa-se a formação de um grande e único bloco de aproximadamente 14 kb em intenso DL, causado pelos sítios de variação rs2675346 (7), rs2433354 (13), rs2433356 (24) rs1426654 (27), os mesmos sítios que desviam do Equilíbrio de Hardy-Weinberg (Tabela 4). Os sítios rs2675346 e rs2433354 apresentaram DL absoluto ($D' = 1$; $r^2 = 1$).

6.7.Associações fenotípicas

Para o estudo de associações os tipos de pele, avaliados durante a coleta de acordo com a escala de Fitzpatrick, foram estratificadas entre três tipos: peles I e II foram classificadas como claras, peles III e IV foram classificadas como intermediárias, e peles V e VI como escuras.

Foi encontrado um total de 285 associações entre alelos ou genótipos de algum sítio de variação analisado e alguma característica ligada à pigmentação dos olhos, cabelos ou pele (incluindo a presença ou ausência de manchas) com relevância estatística, ou seja, valor de probabilidade obtida pelo teste exato de Fisher inferior a 0,05 e *Odds Ratio* superior a 10, com Intervalo de Confiança de 95% não incluindo o valor 1. Tais associações foram encontradas em 13 sítios de variação dentre os 31 analisados, sendo quatro deles na região promotora, oito em íntrons e um em éxon. A Tabela 11 apresenta as associações encontradas com menores valores de probabilidade estimados pelo teste exato de Fisher, $p < 1,000 \times 10^{-15}$. Todas as associações obtidas podem ser encontradas no Apêndice A.

Tabela 11 – Associações entre alelos ou genótipos e fenótipos de pigmentação que demonstraram menores valores de probabilidade por meio do teste exato de Fisher. Estão aqui apresentadas apenas as associações com $p < 1,000 \times 10^{-15}$. O conjunto completo de associações obtidas pode ser encontrado no Apêndice A.

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs2675346 / rs2433354	C/C	Pele clara x pele escura	$2,2900 \times 10^{-19}$	28,3273	64,0585 - 12,5266
rs2675346 / rs2433354	C	Pele clara x pele escura	$2,3100 \times 10^{-24}$	23,3471	47,2028 - 11,5478
rs2675346 / rs2433354	T	Pele escura x pele clara	$2,3100 \times 10^{-24}$	23,3471	47,2028 - 11,5478
rs2675346 / rs2433354	T	Cabelos pretos x cabelos castanhos claros	$3,8700 \times 10^{-16}$	21,4224	60,4283 - 7,5944
rs2675346 / rs2433354	C	Cabelos castanho-claros x cabelos pretos	$3,8700 \times 10^{-16}$	21,4224	60,4283 - 7,5944
rs2433356	G	Pele clara x pele escura	$8,3700 \times 10^{-16}$	25,1647	66,6444 - 9,5021
rs2433356	A	Pele escura x pele clara	$8,3700 \times 10^{-16}$	25,1647	66,6444 - 9,5021
rs1426654	G/G	Pele escura x pele clara	$2,7500 \times 10^{-22}$	214,1538	1634,7776 - 28,0539
rs1426654	A/A	Pele clara x pele escura	$1,5200 \times 10^{-29}$	75,5926	200,5499 - 28,4929
rs1426654	A	Pele clara x pele escura	$1,2200 \times 10^{-46}$	45,7303	84,6774 - 24,6968
rs1426654	G	Pele escura x pele clara	$1,2200 \times 10^{-46}$	45,7303	84,6774 - 24,6968
rs1426654	G	Cabelos pretos x cabelos loiro-escuros	$1,6900 \times 10^{-18}$	33,8608	111,3093 - 10,3006
rs1426654	A	Cabelos loiro-escuros x cabelos pretos	$1,6900 \times 10^{-18}$	33,8608	111,3093 - 10,3006
rs1426654	G/G	Pele escura x não pele escura	$6,1300 \times 10^{-19}$	22,5882	46,0081 - 11,0900
rs1426654	A/A	Cabelos castanho-claros x cabelos pretos	$4,5300 \times 10^{-17}$	21,6000	49,3118 - 9,4614
rs1426654	A/A	Não pele escura x pele escura	$1,4100 \times 10^{-18}$	21,0347	50,6037 - 8,7436
rs1426654	G	Cabelos pretos x cabelos castanho-claros	$4,8500 \times 10^{-24}$	18,4203	37,2770 - 9,1023
rs1426654	A	Cabelos castanho-claros x cabelos pretos	$4,8500 \times 10^{-24}$	18,4203	37,2770 - 9,1023
rs1426654	A/A	Pele clara x não pele clara	$1,7000 \times 10^{-26}$	13,8279	24,2228 - 7,8938
rs1426654	G	Pele escura x não pele escura	$1,2700 \times 10^{-31}$	12,6604	19,9505 - 8,0342
rs1426654	A	Não pele escura x pele escura	$1,2700 \times 10^{-31}$	12,6604	19,9505 - 8,0342
rs1426654	A	Pele clara x não pele clara	$5,9300 \times 10^{-35}$	12,6055	20,8034 - 7,6381
rs1426654	G	Não pele clara x pele clara	$5,9300 \times 10^{-35}$	12,6055	20,8034 - 7,6381

Os SNPs rs2675346 e rs2433354 apresentaram as mesmas 37 associações com parâmetros estatísticos idênticos, o que reflete o DL absoluto encontrado entre eles. Os alelos rs2675346*C e rs2433354*C e seus genótipos C/C estão associados a fenótipos de pigmentação claros de olhos, cabelos (incluindo cabelos ruivos) e pele e fenótipos de pigmentação intermediários de cabelos. Os alelos rs2675346*T e rs2433354*T e seus genótipos T/T e C/T estão associados a fenótipos de pigmentação escuros de pele, cabelos e olhos.

Os sítios de variação rs553733344 e rs55728404 apresentaram as mesmas 27 associações com fenótipos de pigmentação dos olhos, da pele e dos cabelos, entretanto com variações nos valores de intensidade e significância (Apêndice A). Estes sítios também se encontram em alto DL e separados por apenas dez bases nitrogenadas.

O SNP rs1426654 (Ala111Thr) se destacou por apresentar o maior número de associações (46 delas) e as mais fortes. O alelo A (Thr) deste SNP está relacionado à pele clara, enquanto o alelo G (Ala) está relacionado com pele escura. O genótipo G/G está relacionado a peles escuras em comparação com peles claras, com $OR = 214,1538$.

Foi encontrada uma associação entre o genótipo A/G deste SNP com cabelos pretos em comparação com loiro-claros. Outras associações foram obtidas entre o genótipo G/G e o alelo G com fenótipos de pigmentação escuros de cabelos, olhos e pele, e fenótipos de pigmentação intermediários de pele (associação encontrada exclusivamente quando avaliado este sítio) e cabelos (castanho-escuros). O genótipo A/A e alelo A foram associados a fenótipos de pigmentação claros de cabelos, olhos e pele (incluindo cabelos ruivos), além disso, foram também associados a fenótipos intermediários de cabelos (loiro-escuros).

O sítio de variação rs2433356 foi o único a mostrar associações entre o seu genótipo A/A e a ausência de manchas em regiões da pele expostas ao sol (Apêndice A). Este genótipo, o genótipo heterozigoto e o alelo A também foram associados a fenótipos escuros de pele, cabelos e olhos. O genótipo G/G e o alelo G mostraram-se associados a fenótipos claros de pele, cabelos e olhos e fenótipos intermediários de cabelos (loiro-escuros) e olhos (verdes).

Foram encontradas 19 associações significantes entre sete haplótipos e fenótipos de pigmentação de cabelos e pele, incluindo a presença ou ausência de manchas (Tabela 12). Não foram encontradas associações entre haplótipos e fenótipos de pigmentação de olhos.

Tabela 12 – Associações estatisticamente significantes entre sete haplótipos e fenótipos de pigmentação.

<i>HAP</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
HAP 01	Pele clara x pele escura	$4,0978 \times 10^{-35}$	20,1075	33,9074 - 11,9240
HAP 01	Cabelos loiro-escuros x cabelos pretos	$3,1882 \times 10^{-13}$	10,8528	23,0540 - 5,1090
HAP 01	Cabelos ruivos x cabelos pretos	$7,8793 \times 10^{-7}$	29,7273	225,7119 - 3,9152
HAP 01	Cabelos loiro-claros x cabelos pretos	$5,4978 \times 10^{-7}$	12,7403	43,4987 - 3,7315
HAP 02	Pele escura x pele clara	$4,4228 \times 10^{-15}$	18,6667	43,7571 - 7,9632
HAP 02	Cabelos castanho-escuros x cabelos loiro-escuros	$7,9310 \times 10^{-4}$	12,2571	90,6939 - 1,6565
HAP 02	Cabelos pretos x cabelos loiro-escuros	$2,2958 \times 10^{-5}$	18,4800	137,3588 - 2,4863
HAP 03	Pele escura x pele clara	$2,2350 \times 10^{-8}$	47,9020	368,6587 - 6,2242
HAP 03	Pele intermediária x pele clara	$4,9174 \times 10^{-4}$	15,3797	118,2611 - 2,0001
HAP 03	Cabelos pretos x cabelos castanho-claros	$1,2114 \times 10^{-4}$	16,4970	124,7522 - 2,1815
HAP 03	Cabelos pretos x cabelos loiro-escuros	$1,2827 \times 10^{-3}$	18,2776	306,6074 - 1,0896
HAP 04	Cabelos pretos x cabelos castanho-claros	$4,4526 \times 10^{-3}$	10,8960	84,2913 - 1,4085
HAP 05	Pele escura x pele clara	$2,7736 \times 10^{-5}$	16,4151	76,0868 - 3,5414
HAP 06	Pele escura x pele clara	$5,1527 \times 10^{-5}$	48,0137	847,4462 - 2,7203
HAP 06	Cabelos pretos x cabelos castanho-escuros	$1,3826 \times 10^{-3}$	14,3898	114,5516 - 1,8076
HAP 06	Manchas não expostas x manchas expostas	$2,1855 \times 10^{-4}$	13,7674	65,4356 - 2,8966
HAP 07	Cabelos loiro-claros x cabelos castanho-escuros	$4,6964 \times 10^{-2}$	10,0714	74,2747 - 1,3657
HAP 07	Cabelos loiro-claros x cabelos pretos	$1,8734 \times 10^{-2}$	32,7193	699,1105 - 1,5313
HAP 07	Cabelos loiro-claros x cabelos castanho-claros	$2,8247 \times 10^{-2}$	25,7018	549,6717 - 1,2018

Os haplótipos HAP 01 e HAP 03 geraram o maior número de associações, quatro associações cada. A associação entre HAP 01 e peles claras contra peles escuras apresentou o menor valor de probabilidade do teste exato de Fisher ($p = 4,0978 \times 10^{-35}$). Este haplótipo foi ainda responsável por associações com cabelos claros (loiro-escuro e loiro-claro) e a única associação entre haplótipos e cabelos ruivos ($p = 7,8793 \times 10^{-7}$, $OR = 29,7273$). Ele contém os alelos rs2675346*C, rs2433354*C e rs1426654*A, os únicos alelos e genótipos associados a cabelos ruivos, sendo a associação mais intensa a associação entre o alelo rs1426654*A ($p = 9,8800 \times 10^{-7}$, $OR = 28,4430$), que se mostra tão significativa e intensa quanto à associação entre HAP 01 e cabelos ruivos.

A associação entre HAP 03 também com peles escuras contra peles claras apresentou o maior valor de *Odds Ratio* ($OR = 47,9020$). Este haplótipo apresentou ainda associações com cabelos escuros e a única associação entre haplótipos e pele intermediária ($p = 4,9174 \times 10^{-4}$, $OR = 15,3797$). Ele possui o alelo rs1426654*G, sendo o genótipo G/G o único dentre alelos e genótipos associados à pele intermediária ($p = 4,9100 \times 10^{-5}$, $OR = 20,1739$) sendo sua significância dez vezes maior e sua intensidade 25% maior em comparação às encontradas entre HAP 03 e pele intermediária.

O haplótipo HAP 06 foi associado a fenótipos escuros de pele e cabelos (preto) e também foi o único haplótipo associado à ausência de manchas em regiões exposta da pele ($p = 2,1855 \times 10^{-4}$, $OR = 13,7674$). Este haplótipo possui o alelo rs2433356*A, cujo genótipo A/A é o único dentre alelos e genótipo associado à ausência de manchas ($p = 7,8765 \times 10^{-4}$, $OR = 21,1429$), com significância próxima à encontrada entre a ausência de manchas e HAP 06 e intensidade 35% maior.

6.8. Interações TF/TFBS na região promotora do gene *SLC24A5*

A interação entre Fatores de Transcrição (TFs) e *motifs* que incluem sítios de variação associados a fenótipos de pigmentação foi avaliada. Estes *motifs*, também conhecidos como TFBS (sítios de ligação de Fatores de Transcrição) atuam na regulação da expressão do gene *SLC24A5*.

A Tabela 13 apresenta os sítios de variação associados, os TFs que podem ligar-se a TFBS que compreendem estes sítios e o valor de Δ LOD. Este último parâmetro representa a variação da afinidade do TF em relação ao alelo referência ou alternativo do sítio em questão: quando Δ LOD > 0 , o TF possui maior afinidade pelo alelo referência, ou seja, se ligará com maior eficiência quando o alelo referência estiver presente; quando Δ LOD < 0 , o TF possui maior afinidade pelo alelo alternativo, ou seja, se ligará com maior eficiência quando o alelo alternativo estiver presente.

Tabela 13 – Fatores de Transcrição (TFs) cuja eficiência de ligação ao gene *SLC24A5* pode ser alterada pela presença de alelo referência ou alternativo nos sítios de variação associados a algum fenótipo de pigmentação. A afinidade do TF por cada alelo é avaliada pelo valor de Δ LOD.

Sítio de variação	Posição	REF	ALT	Região	Marca de histona	TF	TFBS motifs	Δ LOD
rs10162789	48.410.384	T	G	Promotora	H3K4me1_Enh	Dados não encontrados	AP-2_known5 AP-2_known6 PTF1-beta	-0,8 -1,1 0,4
rs553733344	48.411.795	C	CA	Promotora Região hipersensível à DNase I	H3K4me1_Enh H3K4me3_Pro H3K27ac_Enh	c-FOS, JunD c-FOS, JunD	RREB-1_1 Zfp105	-9,4 -3
rs55728404	48.411.805	T	G	Promotora Região hipersensível à DNase I	H3K4me1_Enh H3K4me3_Pro H3K27ac_Enh	c-FOS, JunD c-FOS, JunD	CIZ SIX5_disc4	-0,8 4,4
rs2675346	48.411.821	C	T	Promotora Região hipersensível à DNase I	H3K4me1_Enh H3K4me3_Pro H3K27ac_Enh	c-FOS, JunD	DMRT5	-10,7
rs2433354	48.414.969	C	T	Íntron	H3K4me3_Pro H3K27ac_Enh	Dados não encontrados	GR_known2	3,8
rs2459391	48.415.068	A	G	Íntron	H3K4me3_Pro H3K27ac_Enh	Dados não encontrados	Foxa_disc2	11,9
rs79856063	48.415.196	T	C	Íntron	H3K4me1_Enh H3K4me3_Pro H3K27ac_Enh	Dados não encontrados	-	-
rs2433355	48.415.261	C	T	Íntron	H3K4me1_Enh H3K4me3_Pro H3K27ac_Enh	Dados não encontrados	Elf5 Ets_known1 Ets_known9	6,3 12 12

Todos os sítios de variação estudados por esta metodologia estão inseridos em uma região com marcas de histonas que ativam a transcrição, H3K4me1_Enh, H3K4me3_Pro e H3K27ac_Enh (Figura 9). Entretanto, apenas os sítios de variação rs553733344, rs55728404 e rs2675346 são alvos de Fatores de Transcrição já identificados (Tabela 13). Estes três sítios estão inseridos em um pequeno segmento de apenas 26 bases, sendo esta uma região hipersensível à DNase I.

Figura 9 – Características de fragmento da região promotora do gene *SLC24A5*, onde se encontram os sítios de variação rs553733344, rs55728404 e rs2675346 (destacados em azul claro), alvos de Fatores de Transcrição (TFs) e associados às marcas de histonas H3K4me1_Enh, H3K4me3_Pro e H3K27ac_Enh. **Fonte:** UCSC Genome Browser.

Os TFs c-FOS e JunD formam o complexo de transcrição AP1, que possui papel de transporte de TYRP1 do endossomo para o melanossomo através da formação de vesículas e túbulos, dirigindo ao favorecimento da produção de eumelanina (Delevoe, et al. 2009). Segundo experimentos de *knockdown* do gene *SLC24A5*, a falta de transcrição deste leva à hipopigmentação (Ginger, et al. 2008). Essas informações levam a crer que quanto maior a afinidade de AP1, maior a transcrição de gene *SLC24A5* e mais intensa será a pigmentação observada.

Para os *motifs* RREB-1_1 e Zfp105 (que incluem o SNP rs55373334), CIZ (que inclui o SNP rs55728404) e DMRT5 (que inclui rs2675346) o valor de Δ LOD indica maior afinidade dos TFs pelos alelos alternativos, que nos três casos estão associados a fenótipos

escuras (pele, olhos e cabelos). Este resultado está de acordo com as informações de função do gene e do complexo de transcrição e das marcas de histonas observadas.

Os haplótipos HAP 04 e HAP 06 possuem os alelos alternativos nos três sítios de variação citados e estão relacionados a cabelos pretos em comparação a castanho-claros (HAP 04) ou castanho-escuros (HAP 06) e pele escura comparada à pele clara (HAP 06). Dessa forma, as associações haplotípicas concordam com as associações alélicas e genotípicas e corroboram os resultados obtidos pela análise *in silico* de Fatores de Transcrição.

6.9. Alterações na sequência primária da proteína NCKX5

O estudo do sítio de variação rs1426654 da região codificante (Ala111Thr) com o software SNPEff e o Ensembl Browser, aponta que, quando o alelo A, que codifica o aminoácido Thr, está presente, ocorre uma perda de função da proteína com risco moderado ao indivíduo. Isso é condizente com as informações anteriormente apresentadas, que indicam que uma proteína NCKX5 defeituosa geraria uma pigmentação menos acentuada em relação à proteína NCKX5 normal. Além disso, esta conclusão é condizente com as associações encontradas por este e outros estudos entre os alelos deste SNP e a intensidade de pigmentação da pele.

6.10. Interações miRNA/mRNA na região 3'UTR do gene *SLC24A5*

A fim de identificar se os miRNAs que se expressam na pele (Joyce, et al. 2011) apresentam afinidade estável e específica pela região 3'UTR do gene *SLC24A5*, foram elencados aqueles que apresentam os maiores valores de estabilidade (E), especificidade (S) e valores combinados de estabilidade e especificidade (ES), de acordo com o software mirhunt. Esse estudo foi realizado apenas considerando a região 3'UTR da sequência de referência do gene *SLC24A5*, visto que não foram encontradas variantes nesta região tanto na amostra populacional de Ribeirão Preto como na amostrada obtida pelo Projeto *1000 Genomes*, o que pode ser indicativo de seleção purificadora. As Tabelas 14, 15 e 16 apresentam os miRNAs

com maiores valores de E, S e ES, respectivamente dentre os que interagem com a região 3'UTR do gene *SLC24A5*. A Tabela 17 sumariza os achados das três tabelas anteriores.

Tabela 14 – Conjunto de 20 miRNAs que apresentam os maiores valores de estabilidade na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*.

miRNA	E
miR-146b-3p	0,7794
miR-346	0,7777
miR-502-5p	0,7543
miR-4732-3p	0,7285
miR-192-3p	0,6941
miR-1247-5p	0,6826
miR-328-3p	0,6751
miR-1913	0,6726
miR-1260a	0,6722
miR-454-5p	0,6719
miR-1976	0,6708
miR-339-5p	0,6697
miR-6509-3p	0,6604
miR-1260b	0,6595
miR-1914-5p	0,6574
miR-10a-5p	0,6401
miR-4725-5p	0,6395
miR-4672	0,6334
miR-19a-5p	0,6333
miR-214-5p	0,6233

Tabela 15 – Conjunto de 20 miRNAs que apresentam os maiores valores de especificidade na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*.

miRNA	S
miR-19a-5p	1,5131
miR-590-3p	1,1838
miR-4520-2-3p	1,0119
miR-449b-3p	1,0006
miR-885-5p	0,9267

miRNA	S
miR-374b-5p	0,8921
miR-454-5p	0,8583
miR-34b-3p	0,8483
miR-192-3p	0,7825
miR-6820-3p	0,7762
miR-146b-3p	0,7719
miR-502-5p	0,7631
miR-6509-3p	0,7618
miR-590-5p	0,7490
miR-153-3p	0,7070
miR-10b-5p	0,6820
miR-1260b	0,6665
miR-10a-5p	0,6627
miR-1260a	0,6468
miR-106a-3p	0,6310

Tabela 16 – Conjunto de 20 miRNAs que apresentam os maiores valores de score combinado de estabilidade e especificidade na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*.

miRNA	ES
miR-19a-5p	4,8106
miR-454-5p	4,0531
miR-590-3p	3,8631
miR-502-5p	3,6523
miR-146b-3p	3,5798
miR-192-3p	3,4799
miR-346	3,3134
miR-6509-3p	3,2964
miR-4732-3p	3,2077
miR-449b-3p	3,1949
miR-1260a	3,1319
miR-1260b	3,1231
miR-1247-5p	3,0786
miR-10a-5p	3,0765
miR-153-3p	3,0148
miR-10b-5p	2,9595
miR-146a-3p	2,9461
miR-4520-2-3p	2,9449
miR-885-5p	2,9411
miR-328-3p	2,8330

Tabela 17 – Conjunto de 32 miRNAs que apresentam alta estabilidade, especificidade e/ou *score* combinado na interação com o haplótipo referência da região 3'UTR do gene *SLC24A5*. Em negrito estão os miRNAs que apresentam altos valores em todos os três parâmetros avaliados.

E	S	ES
	miR-106a-3p	
miR-10a-5p	miR-10a-5p	miR-10a-5p
	miR-10b-5p	miR-10b-5p
miR-1247-5p		miR-1247-5p
miR-1260a	miR-1260a	miR-1260a
miR-1260b	miR-1260b	miR-1260b
		miR-146a-3p
miR-146b-3p	miR-146b-3p	miR-146b-3p
	miR-153-3p	miR-153-3p
miR-1913		
miR-1914-5p		
miR-192-3p	miR-192-3p	miR-192-3p
miR-1976		
miR-19a-5p	miR-19a-5p	miR-19a-5p
miR-214-5p		
miR-328-3p		miR-328-3p
miR-339-5p		
miR-346		miR-346
	miR-34b-3p	
	miR-374b-5p	
	miR-449b-3p	miR-449b-3p
	miR-4520-2-3p	miR-4520-2-3p
miR-454-5p	miR-454-5p	miR-454-5p
miR-4672		
miR-4725-5p		
miR-4732-3p		miR-4732-3p
miR-502-5p	miR-502-5p	miR-502-5p
	miR-590-3p	miR-590-3p
	miR-590-5p	
miR-6509-3p	miR-6509-3p	miR-6509-3p
	miR-6820-3p	
	miR-885-5p	miR-885-5p

Foram encontrados 1.082 miRNAs com interação estável e/ou específica com o haplótipo referência da região 3'UTR do gene *SLC24A5*. Nove miRNAs destacaram-se por apresentarem alta estabilidade, especificidade e valor combinado dos dois parâmetros: miR-10a-5p, miR-1260a, miR-1260b, miR-146b-3p, miR-192-3p, miR-19a-5p, miR-454-5p, miR-502-5p e miR-6509-3p. O miRNA miR-19a-5p se destacou por apresentar o maior valor de S e ES, 1,5131 e 4,8106, respectivamente com alto valor de E, 0,6333.

Visto o pequeno comprimento da região 3'UTR deste gene (apenas 41 nucleotídeos) e a conservação desta região observada nas populações estudadas, os altos valores dos parâmetro avaliados indicam que estes miRNAs podem desempenhar um importante papel na regulação pós-transcricional do gene estudado ou ainda que mutações nos sítios de variação da região 3'UTR sejam eliminadas por gerarem interações com outros miRNAs que alterem a expressão gênica de *SLC24A5*, inibindo a tradução do mRNA a proteína NCKX5.

6.11. Relação entre índice de melanina e fenótipos e genótipos

A fim de encontrar associações entre os 13 sítios de variação associados a fenótipos e os índices de melanina de regiões exposta e pouco exposta ao sol, foram utilizados os testes estatísticos de Mann-Whitney e Kruskal-Wallis, considerando significativas as associações com $p < 0,05$. As médias de índice de melanina e os respectivos desvios padrões para cada genótipo de cada sítio de variação associado a um fenótipo de pigmentação na amostra de Ribeirão Preto, e os valores de p encontrados pelos testes empregados são apresentados na Tabela 18.

Tabela 18 – Médias do índice de melanina relacionadas a genótipos de sítios de variação associados a fenótipos de pigmentação. Valores significativos de p para os testes de Mann-Whitney e Kruskal-Wallis. Em negrito estão os valores de $p > 0,05$, ou seja, não significativo.

Sítio de variação	REF (0)	ALT (1)	Área exposta ao sol							Área pouco exposta ao sol								
			Média do índice de melanina (desvio padrão)			Mann- Whitney	Kruskal-Wallis			Média do índice de melanina (desvio padrão)			Mann- Whitney	Kruskal-Wallis				
			0/0	0/1	1/1	p	p	p 0/0	p 0/0	p 0/1	0/0	0/1	1/1	p	p	p 0/0	p 0/0	p 0/1
								x 0/1	x 1/1	x 1/1						x 0/1	x 1/1	x 1/1
rs10162789	T	G	1,1153 (0,4283)	1,5144 (0,4379)		0,096					1,5174 (0,3784)	1,9642 (0,1753)		0,040				
rs553733344	C	CA	1,0662 (0,3937)	1,6288 (0,4331)		0,000					1,4794 (0,3563)	1,9262 (0,3614)		0,000				
rs55728404	T	G	1,0687 (0,3958)	1,6199 (0,4361)		0,000					1,4814 (0,3578)	1,9184 (0,3636)		0,000				
rs2675346 / rs2433354	C	T	0,9968 (0,3396)	1,4191 (0,4327)	1,9063 (0,2989)		0,000	0,000	0,000	0,018	1,4171 (0,3172)	1,8044 (0,3740)	2,0823 (0,2482)	0,000	0,000	0,000	0,091	
rs2459391	A	G	1,0675 (0,3958)	1,6162 (0,4304)		0,000					1,4794 (0,3563)	1,9260 (0,3612)		0,000				
rs79856063	T	C	1,1153 (0,4283)	1,5144 (0,4379)		0,096					1,5174 (0,3784)	1,9642 (0,1753)		0,040				
rs2433355	C	T		0,6657 (0,2532)	1,1267 (0,4275)	0,002						1,2625 (0,2357)	1,5256 (0,3799)	0,051				
rs79338096	A	C	1,0915 (0,4128)	1,5883 (0,4487)		0,000					1,5028 (0,3688)	1,8366 (0,4268)		0,001				
rs189640907	A	T	1,1185 (0,4312)	1,1084 (0,1625)		0,651					1,5195 (0,3803)	1,6501 (0,2481)		0,318				
rs2433356	G	A	1,0434 (0,3783)	1,5439 (0,4268)	1,9118 (0,3386)		0,000	0,000	0,000	0,603	1,4565 (0,3447)	1,9095 (0,3525)	2,0460 (0,1296)	0,000	0,000	0,000	1,000	
rs138961421	A	C	1,1153 (0,4283)	1,5144 (0,4379)		0,096					1,5174 (0,3784)	1,9642 (0,1753)		0,040				
rs1426654	A	G	0,9271 (0,2897)	1,3103 (0,3609)	1,6834 (0,4749)		0,000	0,000	0,000	0,017	1,3525 (0,2766)	1,7006 (0,3421)	1,9957 (0,3249)	0,000	0,000	0,000	0,004	

Para a região exposta ao sol, apenas as associações entre os sítios rs10162789, rs79856063, rs189640907 e rs138961421 não se mostraram significativas. Para a região pouco exposta ao sol apenas as associações entre os sítios rs2433355 e, novamente, rs189640907 não se apresentaram significativas. Estes sítios não foram associados a fenótipos de pigmentação de pele, além disso, encontram-se em regiões regulatórias e não foram associados a Fatores de Transcrição. Todos os demais sítios, incluindo os sítios de variação rs553733344, rs55728404, rs2675346 (cujas associações com fenótipos de pigmentação foram explicadas pelos TFs) e rs1426654 (com forte associação com fenótipos e cujo alelo A está relacionado à perda de função da proteína), apresentaram associações significativas com o índice de melanina observado nas regiões exposta e pouco exposta ao sol, concordando com as associações encontradas entre seus genótipos e fenótipos de pigmentação de pele.

Discussão

7. Discussão

7.1. Amostra populacional

A amostra populacional obtida na região de Ribeirão Preto se apresentou altamente miscigenada, contendo 67,57% de indivíduos considerados brancos, 17,94% de indivíduos pardos, 11,30% de indivíduos pretos e 3,19% de indivíduos considerados amarelos. Esta proporção pode ser considerada semelhante à população da mesorregião de Ribeirão Preto avaliada pelo censo do IBGE em 2010. Contudo a proporção destes grupos difere em função da região brasileira considerada, sendo sempre a maior proporção de brancos.

A alta miscigenação na população brasileira, originada há mais de 500 anos e devido às diversas migrações, produziu uma grande variedade de fenótipos de pigmentação, especialmente intermediários, que não são facilmente observados em populações com menor intensidade de miscigenação. Deste modo, o impacto da variabilidade genética na determinação de tais fenótipos é menos estudado e compreendido.

A amostra populacional obtida por este estudo apontou inicialmente 65 sítios de variação no gene *SLC24A5*. Contudo, apenas 31 sítios de variação foram de fato considerados, já que os demais sítios consistem em *singletons*, ou seja, com ocorrência de uma única cópia de um dos alelos em toda a amostra populacional. Dos 407 indivíduos amostrados 393 foram genotipados com sucesso. Os demais indivíduos não atingiram o mínimo da qualidade esperada para serem avaliados.

7.2. Desvios do equilíbrio de Hardy-Weinberg e Desequilíbrio de Ligação

Quatro sítios de variação foram encontrados em significativo desvio do Equilíbrio de Hardy-Weinberg e também em forte DL: rs2675346, rs2433354, rs2433356 e rs1426654 (Ala111Thr). Mesmo com a aplicação da correção de Bonferroni para múltiplos testes ($\alpha = 0,05/31 = 0,0016$), apenas o sítio rs2433356 deixaria de apresentar desvio significativo do Equilíbrio de Hardy-Weinberg, e ainda observaríamos três sítios em desvio. Adicionalmente,

a ocorrência de quatro SNPs desviando do equilíbrio não pode ser atribuída ao acaso ($\chi^2 = 4,0764$; $p = 0,0435$). Entretanto, é preciso enfatizar que este cálculo seria mais aplicável a marcadores independentes; devido ao fato de estarem em alto DL, é natural que desvios do Equilíbrio de Hardy-Weinberg sobre alguma região específica, ainda que casuais, afete as regiões vizinhas.

Apesar de ser difícil estabelecer a causa dos desvios observados para os quatro sítios de variação citados, observamos que todos apresentam uma deficiência de heterozigotos, o que é compatível com a ocorrência do efeito Wahlund ou de seleção positiva sobre alguns alelos. A primeira hipótese seria decorrente de uma subestruturação da população amostrada, e se refletiria também em outras regiões genômicas neutras estudadas na mesma amostra populacional, o que não foi observado no estudo de mais de 50 SNPs distribuídos pelo genoma (Valle-Silva, et al. 2019). Para admitir a segunda hipótese, seria esperado, para tais sítios de variação, alguma associação com fenótipos de pigmentação, o que se concretizou nos quatro casos. De fato, observa-se que as associações mais intensas e significativas envolvem os quatro SNPs que se encontram em desequilíbrio de Hardy-Weinberg (Tabela 4) e em forte DL (Figura 8).

A observação do forte DL, chegando a ser perfeito entre os SNPs rs2675346 e rs2433354, foi reconhecida também durante as análises de associações entre fenótipos de pigmentação, onde estes dois sítios apresentaram as mesmas associações com iguais valores de p , *Odds Ratio* e Intervalo de Confiança.

7.3. Associações fenotípicas

Foi encontrado um grande número de associações entre fenótipos de pigmentação e quatro sítios de variação da região promotora e nove da região CDS. Dentre estas associações, estão associações com fenótipos de pigmentação intermediários, como cabelos castanhos, olhos castanhos, olhos verdes, peles de tons III e IV, que não são descritas com grande frequência (Adhikari, et al. 2016; Lock-Andersen, et al. 1998; Park, et al. 2012), mas possuem maior presença em populações miscigenadas. Além destas, foram encontradas associações

com cabelos ruivos, associações menos frequentemente encontradas em nossa população, devido ao pelo baixo número de ruivos aqui encontrados.

Como demonstrado por (Carratto 2017), há uma grande dificuldade de prever fenótipos intermediários utilizando as ferramentas hoje disponíveis, pois estas utilizam combinações entre predições de fenótipos claros e escuros para apontar intermediários, o que gera baixa acurácia nestas predições. O baixo número de associações entre alelos ou genótipos e fenótipos intermediários deve-se à baixa quantidade de estudos com populações miscigenadas, que abrangem um amplo espectro de pigmentação, ou seja, com grande abundância de fenótipos intermediários. Portanto, observamos uma escassez de associações entre genótipos ou alelos e fenótipos intermediários, que devem ser identificadas utilizando populações miscigenadas, como feito no presente estudo. As novas associações descritas podem ser utilizadas em novas ferramentas de predição de fenótipos, gerando maior acurácia na predição de fenótipos intermediários.

Considerando os SNPs organizados em haplótipos, foram encontradas 19 associações entre os sete haplótipos mais frequentes dentre os 39 encontrados em Ribeirão Preto e fenótipos de pigmentação da pele e dos cabelos. O haplótipo HAP 01 mostrou associações altamente significativas com fenótipos de pele e cabelos claros, e foi o único a apresentar associação com cabelos ruivos, com intensidade e significância próxima à encontrada entre cabelos ruivos e o alelo rs1426654*A. O haplótipo HAP 03 mostrou-se associado à pele escura e cabelos pretos, e foi o único a apresentar associações com peles intermediárias, com intensidade e significância menores do que as encontradas com entre pele intermediária e o genótipo G/G de rs1426654. O haplótipo HAP 06 foi associado a peles e cabelos escuros, além da ausência de manchas na pele em regiões expostas ao sol, com significância próxima à encontrada entre a ausência de manchas e o genótipo A/A de rs2433356, entretanto com menor intensidade. Este haplótipo foi responsável pela associação mais forte dentre os haplótipos em função do valor de *OR*.

Giardina, et al. (2008b) avaliou haplótipos deste gene demonstrando seu uso como Marcadores Informativos de Ancestralidade (AIMs, do inglês *Ancestry Informative Markers*), aprimorando estudos anteriores e demonstrando a importância de haplótipos e sua aplicação em estudos populacionais. Entretanto o referido trabalho avaliou haplótipos envolvendo apenas três sítios de variação (rs1426654, rs2555364 e rs16960620) enquanto o presente estudo

avaliou haplótipos envolvendo 31 sítios de variação. Os haplótipos analisados puderam confirmar associações entre alelos ou genótipos e fenótipos de pigmentação com intensidades comparáveis, incluindo associações ainda não descritas, entretanto não agregaram novas associações às já encontradas por alelos e genótipos. Portanto, o esforço de analisar haplótipos, ou seja, o maior gasto computacional e a maior quantidade de sítios a serem avaliados, não se tornou válido.

7.3.1. Região Promotora

Foram encontradas 97 associações distribuídas entre quatro sítios de variação da região promotora, rs10162789, rs553733344, rs55728404 e rs2675346. O sítio de variação rs10162789 foi responsável por apenas quatro das 97 associações entre SNPs da região promotora e fenótipos de pigmentação, sendo todas associações com cabelos pretos. Além disso, estas quatro associações não puderam ser explicadas e, como era esperado, não se mostraram significativas em relação ao índice de melanina na pele, já que não foram encontradas associações com a pigmentação da pele. Já as associações encontradas quanto aos três últimos sítios citados foram explicadas por estudo *in silico* que avaliou a interação de motivos contendo tais sítios de variação com Fatores de Transcrição que promovem uma maior expressão do gene. Adicionalmente estes três sítios encontram-se em região hipersensível à DNase I com marcas de histonas que promovem a sua expressão. As associações destes sítios com o índice de melanina também foram significativas.

Nenhum sítio de variação da região promotora deste gene foi estudado anteriormente e estas são as suas primeiras associações com fenótipos de pigmentação já descritas. Visando sítios de variação que melhor descrevam as características da população brasileira para aplicação em estudos de fenotipagem forense, proporcionando um complemento ao uso do Ala111Thr (rs1426654), observamos três candidatos na região promotora: rs553733344, rs55728404 e rs2675346. As associações a fenótipos de pigmentação encontradas com relação a estes três sítios se tornam importantes por terem sido justificadas pela interação de TFs expressos na pele com a região promotora na qual estão inseridos, provavelmente interferindo

na expressão gênica de *SLC24A5* e minimizando a possibilidade de tais associações terem ocorrido ao acaso.

Entretanto o sítio de variação rs2675346 está em alto DL com rs1426654 (Ala111Thr) (Figura 8), e não agregará muitas informações às já obtidas por Ala111Thr. Por sua vez, os sítios de variação rs553733344 e rs55728404 também apresentam alto DL entre si e não apontam grandes diferenças nas associações obtidas por cada um, já que apresentaram as mesmas 27 associações, com pequenas variações de intensidade e significância. O sítio de variação rs553733344 consiste em uma Indel possuindo inserções A. Já o sítio rs55728404 consiste em uma SNV, sendo o alelo referência T e o alelo alternativo G. O sítio rs553733344 se encontra mais centralizado a região de interação dos TFs, como observado na Figura 9 e possui média de significância e intensidade ligeiramente maiores ($p = 2,5140 \times 10^{-4}$; $OR = 22,3905$) em comparação ao sítio rs55728404 ($p = 3,2239 \times 10^{-4}$; $OR = 22,0210$). Estes fatores apontam que rs553733344 é um melhor candidato a compor estudos de fenotipagem, agregando informações àquelas obtidas por Ala111Thr. Entretanto, o uso de rs55728404 em alternativa a rs553733344 não trará prejuízos.

7.3.2. Região CDS

Para a região CDS do gene *SLC24A5* foram encontradas 188 associações relacionadas a nove sítios de variação. Entretanto apenas as 46 associações envolvendo o sítio de variação rs1426654 (Ala111Thr), o único presente em região exônica, puderam ser explicadas.

As associações apresentadas por Ala111Thr estão de acordo com associações e relatos funcionais já descritos na literatura. A variante 111Thr, originada pelo alelo A, é responsável por uma redução na troca iônica regulada por pH durante a melanogênese (Puri, et al. 2000), sendo frequentemente associada a peles claras, enquanto a variante Ala111, originada pelo alelo G, é relacionada a peles escuras (Dimisianos, et al. 2009; Giardina, et al. 2008b).

Lima et al. (2015) avaliou associações de Ala111Thr e do sítio de variação rs6058017 (*ASIP*) com fenótipos de pigmentação da pele, olhos e cabelos em amostra populacional obtida na cidade de São Paulo. O presente estudo encontrou associações consistentes com

aquelas encontradas para Ala111Thr pelo estudo mencionado, entretanto, aqui obtivemos maior número de associações, além de associações mais fortes.

Sawitzki (2017) estudou a pigmentação da pele e dos olhos de 134 indivíduos da região Sudeste do Brasil e associações com oito sítios de variação pertencentes a sete genes seguindo a metodologia LMC-HMC-SW (do inglês, *Low Melanin Content – High Melanin Content – Snow White Like*) que classifica indivíduos entre fenótipos claros, escuros e combinação de ambos os fenótipos, como pele clara e olhos escuros. O SNP Ala111Thr, único avaliado no gene *SLC24A5*, apontou associações significativas apenas entre o alelo Ala111 e o fenótipo HMC. A metodologia FDA (do inglês, *Factorial Discriminant Analysis*) apontou a formação de clusters contendo 82% dos indivíduos com fenótipos LMC e SW quando o alelo 111Thr foi avaliado.

Durso et al. (2014) analisou associações de 15 sítios de variação, previamente associados à pigmentação, incluindo o haplótipo contendo os sítios rs2555364 e Ala111Thr do gene *SLC24A5*, com a ancestralidade autodeclarada e um índice de pigmentação atribuído de acordo com a probabilidade de possuir maior ancestralidade europeia, africana ou ameríndia. A amostra populacional continha 455 indivíduos das cidades de São Paulo e do Rio de Janeiro. Aos indivíduos foram atribuídos valores numéricos relativos à sua ancestralidade autodeclarada, sendo 0,0 para indivíduos brancos, 0,5 para pardos e 1,0 para negros. Apenas quatro SNPs permaneceram associados ao índice de pigmentação, entre eles o haplótipo contendo os sítios rs2555364 e Ala111Thr. O haplótipo C Ala111, foi associado a pardos e negros, e o haplótipo G 111Thr associado a brancos e pardos. As associações destes quatro SNPs com o índice de pigmentação atribuído foram mais significativas em comparação às associações entre ancestralidade autodeclarada e o índice de pigmentação.

Leite et al. (2011) avaliou associações de 21 SNPs de um painel de AIMs, cor autodeclarada, status socioeconômico, ancestralidade genômica e índice de melanina em amostra de 86 pares de irmãos de Brasília. As ancestralidades médias africana e europeia foram associadas ao índice de melanina, entretanto não houve associação entre a ancestralidade média ameríndia e o índice. O status socioeconômico foi associado apenas à ancestralidade. Dentre os sítios avaliados apenas Ala111Thr foi associado ao índice de melanina, o genótipo homozigoto Ala111/Ala111 foi associado a um aumento no índice de melanina, enquanto o genótipo heterozigoto foi associado à queda no índice de melanina.

Cerqueira et al. (2014) estudou a associação entre 18 SNPs, entre eles Ala111Thr, e o índice de melanina de 352 indivíduos do Rio Grande do Sul e 148 indivíduos da Bahia obtidos pelo consórcio CANDELA. Apenas os sítios Ala111Thr e rs16891982 (*SLC45A2*) mostraram associações significativas com o índice de melanina da amostra total e dividida em função dos estados onde foi obtida, sendo os alelos Ala111 e C, associados a maiores índices de melanina. Os alelos rs1129038*G (*HERC2*) e rs1126809*G (*TYR*) foram associados a maiores índices de melanina apenas na amostra total.

O presente estudo encontrou associações ainda não descritas entre alelos e genótipos deste sítio à pele intermediária e cabelos claros (incluindo ruivos) e escuros. O alelo A e o genótipo A/A foram associados a fenótipos claros de pele, olhos e cabelos e fenótipos intermediários de cabelos. O alelo G e os genótipos A/G (cabelos escuros) e G/G foram associados a fenótipos escuros e intermediários de pele, cabelos e olhos e fenótipos.

Recentemente, este sítio foi aplicado pelo sistema HIrisPlex-S para predição de fenótipos da pele e é amplamente utilizado como marcador de ancestralidade, o que é evidenciado pela clara discrepância das frequências alélicas nas populações africana, leste asiática e europeia (Apêndice B).

Apesar de não ter sido possível explicar as associações encontradas nas regiões intrônicas do gene *SLC24A5*, as associações obtidas para o SNP rs2433354 devem se originar do seu DL absoluto com o sítio de variação da região promotora rs2675346, cujas associações puderam ser explicadas pela ligação de TFs que controlariam a expressão gênica.

7.3.3. Região 3'UTR

A região 3'UTR do gene *SLC24A5* é extremamente pequena, apresentando apenas 41 bases de extensão e 15 sítios de variação descritos até o momento (Figura 10). O fato de esta região não ter apresentado variações na população de Ribeirão Preto e na população obtida pelo Projeto *1000 Genomes* pode apontar que esta região é de alguma forma conservada. miRNAs expressos na pele podem ligar-se a sítios de variação da região 3'UTR de mRNAs e

alterar a expressão gênica do gene *SLC24A5* e, consequentemente, a produção de melanina e a pigmentação na pele do indivíduo (Gulyaeva e Kushlinskiy 2016; Kamaraj, et al. 2014).

Figura 10 – Características de fragmento que compõe a região 3'UTR do gene *SLC24A5*, onde se encontram 18 sítios de variação. **Fonte:** UCSC Genome Browser.

As análises *in silico* de miRNAs apontaram nove miRNAs (miR-10a-5p, miR-1260a, miR-1260b, miR-146b-3p, miR-192-3p, miR-19a-5p, miR-454-5p, miR-502-5p e miR-6509-3p) expressos na pele que se ligariam à região 3'UTR do gene *SLC24A5* com alta especificidade (S) e estabilidade (E), gerando um alto *score* combinado destes parâmetros (ES). Os valores encontrados são compatíveis com outros miRNAs expressos na pele anteriormente descritos como reguladores pós-transcricionais de outros mRNAs relacionados à pigmentação (Fracasso 2018; Marcorin 2017; Oliveira 2016; Pereira 2017).

Recentemente, Vaher, et al. (2019) apontou relação entre o miRNA miR-10a-5p e tecidos afetados por dermatite atópica. Este estudo demonstrou ainda que o miRNA em questão inibe a proliferação de queratinócitos, o que poderia afetar a distribuição de melanina. Latchana, et al. (2016) avaliou a presença de miRNAs em lesões melanocíticas de Spitz e apontou maior expressão do miRNA miR-1260a em células de tumores benignos de Spitz. Avaliando miRNAs em células de carcinomas cutâneos, (Sand, et al. (2017)) encontrou maior expressão do miRNA miR-19a-5p em carcinoma de células escamosas cutâneas. Este miRNA se destacou por apresentar o maior valor de especificidade e o maior *score* combinado, possuindo também alto valor de estabilidade. A expressão dos outros seis miRNAs na pele foi relatada apenas por (Joyce, et al. (2011)) até o momento. Os miRNAs descritos são candidatos a desempenhar papéis como reguladores pós-transcricionais do gene *SLC24A5*.

Conclusões

8. Conclusões

Foram encontrados 65 sítios de variação no gene *SLC24A5* na amostra populacional de Ribeirão Preto, entretanto apenas 31 não são *singletons* e foram devidamente avaliados, nove na região promotora e 22 na região CDS. Dentre os 31 sítios avaliados, nove são polimórficos. Quatro deles (rs2675346, rs2433354, rs2433356 e rs1426654 (Ala111Thr)) desviam significativamente das proporções esperadas sob Equilíbrio de Hardy-Weinberg e se apresentam em forte DL.

Sete haplótipos dentre os 39 encontrados na população de Ribeirão Preto apresentaram associações com pelo menos um dos fenótipos de pigmentação estudados. Estas associações apenas confirmam as associações encontradas para alelos e genótipos e se mostram comparáveis ou menores em termos de intensidade e significância. Portanto, o esforço de se utilizar haplótipos (maior gasto computacional e maior quantidade de sítios a serem avaliados) não é recompensado.

Dentre as 285 associações significativas encontradas, 139 foram biologicamente explicadas de acordo com a interação de Fatores de Transcrição que alteram a expressão do gene *SLC24A5* ou pelo papel desempenhado pela mutação na proteína NCKX5 produzida. Foram encontradas associações com fenótipos de pigmentação intermediários de pele, cabelos e olhos, associações de grande importância em populações tão miscigenadas quanto à brasileira. Adicionalmente, foram encontradas associações com cabelos ruivos, associações pouco encontradas na população brasileira devido ao baixo número de indivíduos que apresentam tal fenótipo.

Três sítios de variação da região promotora, rs553733344, rs55728404 e rs2675346, apresentaram associações significativas e explicadas pelos TFs c-FOS e JunD. Estes sítios não haviam sido relacionados a fenótipos de pigmentação anteriormente e apontam a grande importância da variabilidade genética da região promotora deste gene. O SNP rs2675346 está em DL absoluto com o SNP rs2433354, o que nos levou a encontrar as mesmas associações para ambos os sítios. Os sítios rs553733344 e rs55728404 também apresentam alto DL entre si, com as mesmas associações, porém valores de intensidade e significância ligeiramente distintos.

O SNP Ala111Thr apresentou as mais fortes associações dentre as obtidas. Este SNP é amplamente conhecido e já é utilizado como marcador de ancestralidade e de predição, como, por exemplo, pelo sistema HIrisPlex-S. Aqui foram obtidas associações ainda não descritas entre alelos e genótipos deste sítio e fenótipos de cabelos e pele intermediária.

Sugere-se que o sítio de variação rs553733344 da região promotora que apresentou associações com fenótipos de pigmentação explicados pela interação com TFs possa ser utilizado como preditor de fenótipos para aplicação em Fenotipagem de DNA Forense em conjunto com o sítio de variação Ala111Thr.

Referências

- Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacon-Duque JC, Al-Saadi F, Johansson JA, Quinto-Sanchez M, Acuna-Alonzo V, Jaramillo C, Arias W, Barquera Lozano R, Macin Perez G, Gomez-Valdes J, Villamil-Ramirez H, Hunemeier T, Ramallo V, Silva de Cerqueira CC, Hurtado M, Villegas V, Granja V, Gallo C, Poletti G, Schuler-Faccini L, Salzano FM, Bortolini MC, Canizales-Quinteros S, Rothhammer F, Bedoya G, Gonzalez-Jose R, Headon D, Lopez-Otin C, Tobin DJ, Balding D, Ruiz-Linares A 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nat Commun* 7: 10815.
- Ainger SA, Jagirdar K, Lee KJ, Soyer HP, Sturm RA 2017. Skin Pigmentation Genetics for the Clinic. *Dermatology* 233: 1-15.
- Ancans J, Tobin DJ, Hoogduijn MJ, Smit NP, Wakamatsu K, Thody AJ 2001. Melanosomal pH controls rate of melanogenesis, eumelanin/phaeomelanin ratio and melanosome maturation in melanocytes and melanoma cells. *Exp Cell Res* 268: 26-35.
- Barrett JC, Fry B, Maller J, Daly MJ 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263-265.
- Basu Mallick C, Iliescu FM, Mols M, Hill S, Tamang R, Chaubey G, Goto R, Ho SY, Gallego Romero I, Crivellaro F, Hudjashov G, Rai N, Metspalu M, Mascie-Taylor CG, Pitchappan R, Singh L, Mirazon-Lahr M, Thangaraj K, Vilems R, Kivisild T 2013. The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet* 9: e1003912.
- Bertolotti A, Lasseaux E, Plaisant C, Trimouille A, Morice-Picard F, Rooryck C, Lacombe D, Couppie P, Arveiler B 2016. Identification of a homozygous mutation of SLC24A5 (OCA6) in two patients with oculocutaneous albinism from French Guiana. *Pigment Cell Melanoma Res* 29: 104-106.
- Bonilla C, Ness AR, Wills AK, Lawlor DA, Lewis SJ, Davey Smith G 2014. Skin pigmentation, sun exposure and vitamin D levels in children of the Avon Longitudinal Study of Parents and Children. *BMC Public Health* 14: 597.
- Busch A, Richter AS, Backofen R 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* 24: 2849-2856.
- Caliebe A, Walsh S, Liu F, Kayser M, Krawczak M 2017. Likelihood ratio and posterior odds in forensic genetics: Two sides of the same coin. *Forensic Sci Int Genet* 28: 203-210.
- Canfield VA, Berg A, Peckins S, Wentzel SM, Ang KC, Oppenheimer S, Cheng KC 2013. Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3 (Bethesda)* 3: 2059-2067.
- Carratto TMTS, Castelli, EC, Mendes-Junior, CT 2017. Análise dos marcadores do sistema HIRISplex em amostra da população brasileira utilizando sequenciamento de nova geração. *Simpósio Internacional de Iniciação Científica da USP; Ribeirão Preto*.
- Cartegni L, Chew SL, Krainer AR 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet* 3: 285-298.

Castelli EC, Mendes-Junior CT, Sabbagh A, Porto IO, Garcia A, Ramalho J, Lima TH, Massaro JD, Dias FC, Collares CV, Jamonneau V, Bucheton B, Camara M, Donadi EA 2015. HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Hum Immunol* 76: 945-953.

Cerqueira CC, Hunemeier T, Gomez-Valdes J, Ramallo V, Volasko-Krause CD, Barbosa AA, Vargas-Pinilla P, Dornelles RC, Longo D, Rothhammer F, Bedoya G, Canizales-Quinteros S, Acuna-Alonzo V, Gallo C, Poletti G, Gonzalez-Jose R, Salzano FM, Callegari-Jacques SM, Schuler-Faccini L, Ruiz-Linares A, Catira Bortolini M, for C 2014. Implications of the admixture process in skin color molecular assessment. *PLoS One* 9: e96886.

Chaitanya L, Breslin K, Zuniga S, Wirken L, Pospiech E, Kukla-Bartoszek M, Sijen T, Knijff P, Liu F, Branicki W, Kayser M, Walsh S 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.

Cichorek M, Wachulska M, Stasiewicz A, Tyminska A 2013. Skin melanocytes: biology and development. *Postepy Dermatol Alergol* 30: 30-41.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6: 80-92.

Delevoye C, Hurbain I, Tenza D, Sibarita JB, Uzan-Gafsou S, Ohno H, Geerts WJ, Verkleij AJ, Salamero J, Marks MS, Raposo G 2009. AP-1 and KIF13A coordinate endosomal sorting and positioning during melanosome biogenesis. *J Cell Biol* 187: 247-264.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43: 491-498.

Dimisianos G, Stefanaki I, Nicolaou V, Sypsa V, Antoniou C, Poulou M, Papadopoulos O, Gogas H, Kanavakis E, Nicolaidou E, Katsambas AD, Stratigos AJ 2009. A study of a single variant allele (rs1426654) of the pigmentation-related gene SLC24A5 in Greek subjects. *Exp Dermatol* 18: 175-177.

Durso DF, Bydlowski SP, Hutz MH, Suarez-Kurtz G, Magalhaes TR, Pena SD 2014. Association of genetic variants with self-assessed color categories in Brazilians. *PLoS One* 9: e83926.

Excoffier L, Lischer HE 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10: 564-567.

Fitzpatrick, TB 1988. The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol* 124: 869-871.

Fracasso NCA 2018. Diversidade genética das regiões regulatórias e codificantes dos genes SLC45A2 e TYR em amostra da população brasileira. [[Ribeirão Preto]: Universidade de São Paulo.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper

R, Ward R, Lander ES, Daly MJ, Altshuler D 2002. The structure of haplotype blocks in the human genome. *Science* 296: 2225-2229.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR 2015. A global reference for human genetic variation. *Nature* 526: 68-74.

Giardina E, Pietrangeli I, Martinez-Labarga C, Martone C, de Angelis F, De Stefano G, Rickards O, Novelli G 2008a. Haplotypes in SLC24A5 gene as ancestry informative markers in different populations. *Current Genomics* 9: 110-114. doi: Doi 10.2174/138920208784139528

Giardina E, Pietrangeli I, Martinez-Labarga C, Martone C, de Angelis F, Spinella A, De Stefano G, Rickards O, Novelli G 2008b. Haplotypes in SLC24A5 Gene as Ancestry Informative Markers in Different Populations. *Curr Genomics* 9: 110-114.

Ginger RS, Askew SE, Ogborne RM, Wilson S, Ferdinando D, Dadd T, Smith AM, Kazi S, Szerencsei RT, Winkfein RJ, Schnetkamp PP, Green MR 2008. SLC24A5 encodes a trans-Golgi network protein with potassium-dependent sodium-calcium exchange activity that regulates human epidermal melanogenesis. *J Biol Chem* 283: 5486-5495.

Gulyaeva LF, Kushlinskiy NE 2016. Regulatory mechanisms of microRNA expression. *J Transl Med* 14: 143.

Hartl DL, Clark, A. G. 2010. *Princípios de genética de populações* Porto Alegre.

Hearing VJ 2011. Determination of melanin synthetic pathways. *J Invest Dermatol* 131: E8-E11.

Henriksen T, Dahlback A, Larsen SH, Moan J 1990. Ultraviolet-radiation and skin cancer. Effect of an ozone layer depletion. *Photochem Photobiol* 51: 579-582.

IBM C. 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

Jablonski NG, Chaplin G 2010. Colloquium paper: human skin pigmentation as an adaptation to UV radiation. *Proc Natl Acad Sci U S A* 107 Suppl 2: 8962-8968.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS 2004. Human MicroRNA targets. *PLoS Biol* 2: e363.

Joyce CE, Zhou X, Xia J, Ryan C, Thrash B, Menter A, Zhang W, Bowcock AM 2011. Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. *Hum Mol Genet* 20: 4025-4040.

Kamaraj B, Gopalakrishnan C, Purohit R 2014. In silico analysis of miRNA-mediated gene regulation in OCA and OA genes. *Cell Biochem Biophys* 70: 1923-1932.

Kawakami A, Fisher DE 2011. Key discoveries in melanocyte development. *J Invest Dermatol* 131: E2-4.

Kim KI, Jeong HB, Ro H, Lee JH, Kim CD, Yoon TJ 2017. Inhibitory effect of 5-iodotubercidin on pigmentation. *Biochem Biophys Res Commun* 490: 1282-1286.

Kraev A, Quednau BD, Leach S, Li XF, Dong H, Winkfein R, Perizzolo M, Cai X, Yang R, Philipson KD, Lytton J 2001. Molecular cloning of a third member of the potassium-dependent sodium-calcium exchanger gene family, NCKX3. *J Biol Chem* 276: 23161-23172.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynech MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'Donnell D, Kittles R, Parra EJ, Mangini

NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310: 1782-1786.

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haq S, Haq S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietiläinen KH, Pouta A, Ridderstrale M, Rotter JJ, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeny L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley TH, Jr., Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllenstein U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Volzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467: 832-838.

Latchana N, Regan K, Howard JH, Aldrink JH, Ranalli MA, Peters SB, Zhang X, Gru A, Payne PRO, Suarez-Kelly LP, Carson WE, 3rd 2016. Global microRNA profiling for diagnostic appraisal of melanocytic Spitz tumors. *J Surg Res* 205: 350-358.

Leite TK, Fonseca RM, de Franca NM, Parra EJ, Pereira RW 2011. Genomic ancestry, self-reported "color" and quantitative measures of skin pigmentation in Brazilian admixed siblings. *PLoS One* 6: e27162.

Li H, Durbin R 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

Li XF, Kiedrowski L, Tremblay F, Fernandez FR, Perizzolo M, Winkfein RJ, Turner RW, Bains JS, Rancourt DE, Lytton J 2006. Importance of K⁺-dependent Na⁺/Ca²⁺-exchanger 2, NCKX2, in motor learning and memory. *J Biol Chem* 281: 6273-6282.

Li XF, Kraev AS, Lytton J 2002. Molecular cloning of a fourth member of the potassium-dependent sodium-calcium exchanger gene family, NCKX4. *J Biol Chem* 277: 48410-48417.

Li XF, Lytton J 2014. An essential role for the K⁺-dependent Na⁺/Ca²⁺-exchanger, NCKX4, in melanocortin-4-receptor-dependent satiety. *J Biol Chem* 289: 25445-25459.

Lima FA, Goncalves FT, Fridman C 2015. SLC24A5 and ASIP as phenotypic predictors in Brazilian population for forensic purposes. *Leg Med (Tokyo)* 17: 261-266.

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: 841-853.

Lin JY, Fisher DE 2007. Melanocyte biology and skin pigmentation. *Nature* 445: 843-850.

Liu F, Wen B, Kayser M 2013. Colorful DNA polymorphisms in humans. *Semin Cell Dev Biol* 24: 562-575.

Liu F, Wollstein A, Hysi PG, Ankra-Badu GA, Spector TD, Park D, Zhu G, Larsson M, Duffy DL, Montgomery GW, Mackey DA, Walsh S, Lao O, Hofman A, Rivadeneira F, Vingerling JR, Uitterlinden AG, Martin NG, Hammond CJ, Kayser M 2010a. Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet* 6: e1000934.

Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S, Investigators A 2010b. A Versatile Gene-Based Test for Genome-wide Association Studies. *American Journal of Human Genetics* 87: 139-145.

Lock-Andersen J, Wulf HC, Knudstorp ND 1998. Interdependence of eye and hair colour, skin type and skin pigmentation in a Caucasian population. *Acta Derm Venereol* 78: 214-219.

Maddodi N, Jayanthi A, Setaluri V 2012. Shining light on skin pigmentation: the darker and the brighter side of effects of UV radiation. *Photochem Photobiol* 88: 1075-1082.

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.

Martin M 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytisky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20: 1297-1303.

Miller SA, Dykes DD, Polesky HF 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 16: 1215.

Morice-Picard F, Lasseaux E, Francois S, Simon D, Rooryck C, Bieth E, Colin E, Bonneau D, Journel H, Walraedt S, Leroy BP, Meire F, Lacombe D, Arveiler B 2014. SLC24A5 mutations are associated with non-syndromic oculocutaneous albinism. *J Invest Dermatol* 134: 568-571.

Musgrave-Brown E, Ballard D, Balogh K, Bender K, Berger B, Bogus M, Borsting C, Brion M, Fondevila M, Harrison C, Oguzturun C, Parson W, Phillips C, Proff C, Ramos-Luis E, Sanchez JJ, Sanchez Diz P, Sobrino Rey B, Stradmann-Bellinghausen B, Thacker C, Carracedo A, Morling N, Scheithauer R, Schneider PM, Syndercombe Court D 2007. Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet* 1: 186-190.

Nan HM, Kraft P, Hunter DJ, Han JL 2009. Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. *Int J Cancer* 125: 909-917.

Oliveira MLG 2016. Sequenciamento de nova geração do gene IRF4: identificação de variações associadas a fenótipos de pigmentação na população brasileira. [[Ribeirão Preto]: Universidade de São Paulo.

Park SL, Le Marchand L, Wilkens LR, Kolonel LN, Henderson BE, Zhang ZF, Setiawan VW 2012. Risk factors for malignant melanoma in white and non-white/non-African American populations: the multiethnic cohort. *Cancer Prev Res (Phila)* 5: 423-434.

Patarra NLF, Fernandes D 2011. Brasil: país de imigração? *Revista Internacional em Língua Portuguesa* 24: 65-96.

Pena SD, Di Pietro G, Fuchshuber-Moraes M, Genro JP, Hutz MH, Kehdy Fde S, Kohlrausch F, Magno LA, Montenegro RC, Moraes MO, de Moraes ME, de Moraes MR, Ojopi EB, Perini JA, Racciopi C, Ribeiro-Dos-Santos AK, Rios-Santos F, Romano-Silva MA, Sortica VA, Suarez-Kurtz G 2011. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* 6: e17063.

Pereira ALE 2017. Diversidade das regiões regulatórias e exônicas dos genes ASIP, MC1R e TYRP1 determinada por sequenciamento de nova geração em amostra da população brasileira e seu envolvimento na biossíntese de melanina [[Ribeirão Preto]: Universidade de São Paulo.

Porto IO, Mendes-Junior CT, Felicio LP, Georg RC, Moreau P, Donadi EA, Chies JA, Castelli EC 2015. MicroRNAs targeting the immunomodulatory HLA-G gene: a new survey searching for microRNAs with potential to regulate HLA-G. *Mol Immunol* 65: 230-241.

Pospiech E, Chen Y, Kukla-Bartoszek M, Breslin K, Aliferi A, Andersen JD, Ballard D, Chaitanya L, Freire-Aradas A, van der Gaag KJ, Giron-Santamaria L, Gross TE, Gysi M, Huber G, Mosquera-Miguel A, Muralidharan C, Skowron M, Carracedo A,

Haas C, Morling N, Parson W, Phillips C, Schneider PM, Sijen T, Syndercombe-Court D, Vennemann M, Wu S, Xu S, Jin L, Wang S, Zhu G, Martin NG, Medland SE, Branicki W, Walsh S, Liu F, Kayser M, Consortium EU-N 2018. Towards broadening Forensic DNA Phenotyping beyond pigmentation: Improving the prediction of head hair shape from DNA. *Forensic Sci Int Genet* 37: 241-251.

Pospiech E, Karlowska-Pik J, Marcinska M, Abidi S, Andersen JD, Berge MVD, Carracedo A, Eduardoff M, Freire-Aradas A, Morling N, Sijen T, Skowron M, Sochtig J, Syndercombe-Court D, Weiler N, Schneider PM, Ballard D, Borsting C, Parson W, Phillips C, Branicki W, Consortium EU-N 2015. Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. *Forensic Sci Int Genet* 19: 280-288.

Prinsen CF, Szerencsei RT, Schnetkamp PP 2000. Molecular cloning and functional expression of the potassium-dependent sodium-calcium exchanger from human and chicken retinal cone photoreceptors. *J Neurosci* 20: 1424-1434.

Puri N, Gardner JM, Brilliant MH 2000. Aberrant pH of melanosomes in pink-eyed dilution (p) mutant melanocytes. *J Invest Dermatol* 115: 607-613.

Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R 2004. Fast and effective prediction of microRNA/target duplexes. *RNA* 10: 1507-1517.

RIBPG. 2019. X Relatório da Rede Integrada de Bancos de Perfis Genéticos In. Brasília Comitê gestor RIBPG.

Richardson K, Lai CQ, Parnell LD, Lee YC, Ordovas JM 2011. A genome-wide survey for SNPs altering microRNA seed sites identifies functional candidates in GWAS. *BMC Genomics* 12: 504.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP 2011. Integrative genomics viewer. *Nat Biotechnol* 29: 24-26.

Rossberg W, Saturnus R, Wagenpfeil S, Kleber M, Marz W, Reichrath S, Vogt T, Reichrath J 2016. Human Pigmentation, Cutaneous Vitamin D Synthesis and Evolution: Variants of Genes (SNPs) Involved in Skin Pigmentation Are Associated with 25(OH)D Serum Concentration. *Anticancer Res* 36: 1429-1437.

RStudio T. 2015. RStudio: Integrated Development for R. . RStudio, Inc., Boston, MA

Sand M, Hessam S, Amur S, Skrygan M, Bromba M, Stockfleth E, Gambichler T, Bechara FG 2017. Expression of oncogenic miR-17-92 and tumor suppressive miR-143-145 clusters in basal cell carcinoma and cutaneous squamous cell carcinoma. *J Dermatol Sci* 86: 142-148.

Sawitzki FR, Rodenbusch R, Gubert DW, Silva DSBS, Silva EFA, Alho CS 2017. Analysis of Eight SNPs in South Brazilian Subjects with Different Skin and Eye Melanin Content. *SM Journal of Forensic Research and Criminology* 1: 9.

Schnetkamp PP, Szerencsei RT, Basu DK 1991. Unidirectional Na⁺, Ca²⁺, and K⁺ fluxes through the bovine rod outer segment Na-Ca-K exchanger. *J Biol Chem* 266: 198-206.

Stephan AB, Tobochnik S, Dibattista M, Wall CM, Reisert J, Zhao H 2011. The Na(+)/Ca(2+) exchanger NCKX4 governs termination and adaptation of the mammalian olfactory response. *Nat Neurosci* 15: 131-137.

Stephens M, Donnelly P 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73: 1162-1169.

Stephens M, Smith NJ, Donnelly P 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68: 978-989.

Sturm RA, Box NF, Ramsay M 1998. Human pigmentation genetics: the difference is only skin deep. *Bioessays* 20: 712-721.

Sturm RA, Teasdale RD, Box NF 2001. Human pigmentation genes: identification, structure and consequences of polymorphic variation. *Gene* 277: 49-62.

Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Palsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemenev LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K 2007. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 39: 1443-1452.

Thorvaldsdottir H, Robinson JT, Mesirov JP 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14: 178-192.

Tiosano D, Audi L, Climer S, Zhang WX, Templeton AR, Fernandez-Cancio M, Gershoni-Baruch R, Sanchez-Muro JM, El Kholy M, Hochberg Z 2016. Latitudinal Clines of the Human Vitamin D Receptor and Skin Color Genes. *G3-Genes Genomes Genetics* 6: 1251-1266.

Vaher H, Runnel T, Urgard E, Aab A, Carreras Badosa G, Maslovskaja J, Abram K, Raam L, Kaldvee B, Annilo T, Tkaczyk ER, Maimets T, Akdis CA, Kingo K, Rebane A 2019. miR-10a-5p is increased in atopic dermatitis and has capacity to inhibit keratinocyte proliferation. *Allergy*.

Valle-Silva GD, Souza FDN, Marcorin L, Pereira ALE, Carratto TMT, Debortoli G, Oliveira MLG, Fracasso NCA, Andrade ES, Donadi EA, Norton HL, Parra EJ, Simoes AL, Castelli EC, Mendes-Junior CT 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.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43: 11 10 11-33.

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: 170-180.

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: 98-115.

Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Seland JH, Soubrane G, Tomazzoli L, Topouzis F, Vingerling JR, Vioque J, Fletcher AE, Ballantyne KN, Kayser M 2012. DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci Int Genet* 6: 330-340.

Ward LD, Kellis M 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40: D930-934.

Wilson S, Ginger RS, Dadd T, Gunn D, Lim FL, Sawicka M, Sandel M, Schnetkamp PP, Green MR 2013. NCKX5, a natural regulator of human skin colour variation, regulates the expression of key pigment genes MC1R and alpha-MSH and alters cholesterol homeostasis in normal human melanocytes. *Adv Exp Med Biol* 961: 95-107.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newman V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P 2018. Ensembl 2018. *Nucleic Acids Res* 46: D754-D761.

Apêndice

Apêndice A – Tabela 11 (completa) – Associações entre alelos ou genótipos e fenótipos de pigmentação.

Tabela 11 (completa) – Associações entre alelos ou genótipos e fenótipos de pigmentação que demonstraram menores valores de probabilidade por meio do teste exato de Fisher encontradas entre as 285 associações envolvendo 13 sítios de variação.

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs10162789	T/G	Cabelos pretos x cabelos não pretos	$1,3538 \times 10^{-2}$	22,7790	445,1338 - 1,1657
rs10162789	T/T	Cabelos não pretos x cabelos pretos	$1,3538 \times 10^{-2}$	22,7790	445,1338 - 1,1657
rs10162789	G	Cabelos pretos x cabelos não pretos	$1,3708 \times 10^{-2}$	22,4496	436,6315 - 1,1543
rs10162789	T	Cabelos não pretos x cabelos pretos	$1,3708 \times 10^{-2}$	22,4496	436,6315 - 1,1543
rs553733344	C/C	Olhos verdes x olhos castanho-escuros	$1,0500 \times 10^{-5}$	32,8545	544,5689 - 1,9822
rs553733344	C/CA	Olhos castanho-escuros x olhos verdes	$1,0500 \times 10^{-5}$	32,8545	544,5689 - 1,9822
rs553733344	C/CA	Cabelos pretos x cabelos loiro-escuros	$5,5700 \times 10^{-5}$	31,0148	523,0702 - 1,8390
rs553733344	C/C	Cabelos loiro-escuros x cabelos pretos	$5,5700 \times 10^{-5}$	31,0148	523,0702 - 1,8390
rs553733344	C	Olhos verdes x olhos castanho-escuros	$2,0300 \times 10^{-5}$	29,3719	483,1307 - 1,7857
rs553733344	CA	Olhos castanho-escuros x olhos verdes	$2,0300 \times 10^{-5}$	29,3719	483,1307 - 1,7857
rs553733344	C/CA	Cabelos pretos x cabelos castanho-claros	$8,8100 \times 10^{-7}$	27,9403	211,6407 - 3,6886
rs553733344	C/C	Cabelos castanho-claros x cabelos pretos	$8,8100 \times 10^{-7}$	27,9403	211,6407 - 3,6886
rs553733344	CA	Cabelos pretos x cabelos loiro-escuros	$1,3475 \times 10^{-4}$	25,9221	430,9205 - 1,5594
rs553733344	C	Cabelos loiro-escuros x cabelos pretos	$1,3475 \times 10^{-4}$	25,9221	430,9205 - 1,5594
rs553733344	C/C	Pele clara x pele escura	$1,0200 \times 10^{-9}$	25,8000	91,8473 - 7,2472
rs553733344	C/CA	Pele escura x pele clara	$1,0200 \times 10^{-9}$	25,8000	91,8473 - 7,2472
rs553733344	CA	Cabelos pretos x cabelos castanho-claros	$2,2800 \times 10^{-6}$	23,5625	175,8373 - 3,1574
rs553733344	C	Cabelos castanho-claros x cabelos pretos	$2,2800 \times 10^{-6}$	23,5625	175,8373 - 3,1574

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs553733344	C	Pele clara x pele escura	$3,0900 \times 10^{-9}$	21,2449	73,6087 - 6,1317
rs553733344	CA	Pele escura x pele clara	$3,0900 \times 10^{-9}$	21,2449	73,6087 - 6,1317
rs553733344	C/CA	Olhos castanho-escuros x olhos azuis	$1,0332 \times 10^{-3}$	19,6182	327,3382 - 1,1758
rs553733344	C/C	Olhos azuis x olhos castanho-escuros	$1,0332 \times 10^{-3}$	19,6182	327,3382 - 1,1758
rs553733344	C/C	Olhos verdes x olhos não verdes	$5,6292 \times 10^{-4}$	19,0784	314,3920 - 1,1577
rs553733344	C/CA	Olhos não verdes x olhos verdes	$5,6292 \times 10^{-4}$	19,0784	314,3920 - 1,1577
rs553733344	C	Olhos verdes x olhos não verdes	$7,5782 \times 10^{-4}$	17,8187	291,8127 - 1,0880
rs553733344	CA	Olhos não verdes x olhos verdes	$7,5782 \times 10^{-4}$	17,8187	291,8127 - 1,0880
rs553733344	CA	Olhos castanho-escuros x olhos azuis	$9,4060 \times 10^{-4}$	17,4959	288,7539 - 1,0601
rs553733344	C	Olhos azuis x olhos castanho-escuros	$9,4060 \times 10^{-4}$	17,4959	288,7539 - 1,0601
rs553733344	C/C	Pele clara x pele não clara	$2,2800 \times 10^{-7}$	11,3371	37,5508 - 3,4228
rs553733344	C/CA	Pele não clara x pele clara	$2,2800 \times 10^{-7}$	11,3371	37,5508 - 3,4228
rs553733344	C	Pele clara x pele não clara	$4,2000 \times 10^{-7}$	10,4070	34,1381 - 3,1726
rs553733344	CA	Pele não clara x pele clara	$4,2000 \times 10^{-7}$	10,4070	34,1381 - 3,1726
rs55728404	T/T	Olhos verdes x olhos castanho-escuros	$1,0500 \times 10^{-5}$	32,8545	544,5689 - 1,9822
rs55728404	T/G	Olhos castanho-escuros x olhos verdes	$1,0500 \times 10^{-5}$	32,8545	544,5689 - 1,9822
rs55728404	T/G	Cabelos pretos x cabelos loiro-escuros	$5,5700 \times 10^{-5}$	31,0148	523,0702 - 1,8390
rs55728404	T/T	Cabelos loiro-escuros x cabelos pretos	$5,5700 \times 10^{-5}$	31,0148	523,0702 - 1,8390
rs55728404	T	Olhos verdes x olhos castanho-escuros	$2,0300 \times 10^{-5}$	29,3719	483,1307 - 1,7857
rs55728404	G	Olhos castanho-escuros x olhos verdes	$2,0300 \times 10^{-5}$	29,3719	483,1307 - 1,7857
rs55728404	T/G	Cabelos pretos x cabelos castanho-claros	$8,8100 \times 10^{-7}$	27,9403	211,6407 - 3,6886
rs55728404	T/T	Cabelos castanho-claros x cabelos pretos	$8,8100 \times 10^{-7}$	27,9403	211,6407 - 3,6886
rs55728404	G	Cabelos pretos x cabelos loiro-escuros	$1,3475 \times 10^{-4}$	25,9221	430,9205 - 1,5594
rs55728404	T	Cabelos loiro-escuros x cabelos pretos	$1,3475 \times 10^{-4}$	25,9221	430,9205 - 1,5594

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs55728404	T/T	Pele clara x pele escura	$4,5700 \times 10^{-9}$	23,7724	84,9693 - 6,6509
rs55728404	T/G	Pele escura x pele clara	$4,5700 \times 10^{-9}$	23,7724	84,9693 - 6,6509
rs55728404	G	Cabelos pretos x cabelos castanho-claros	$2,2800 \times 10^{-6}$	23,5625	175,8373 - 3,1574
rs55728404	T	Cabelos castanho-claros x cabelos pretos	$2,2800 \times 10^{-6}$	23,5625	175,8373 - 3,1574
rs55728404	T	Pele clara x pele escura	$1,2000 \times 10^{-8}$	19,8620	69,1539 - 5,7046
rs55728404	G	Pele escura x pele clara	$1,2000 \times 10^{-8}$	19,8620	69,1539 - 5,7046
rs55728404	T/G	Olhos castanho-escuros x olhos azuis	$1,0332 \times 10^{-3}$	19,6182	327,3382 - 1,1758
rs55728404	T/T	Olhos azuis x olhos castanho-escuros	$1,0332 \times 10^{-3}$	19,6182	327,3382 - 1,1758
rs55728404	T/T	Olhos verdes x olhos não verdes	$9,9088 \times 10^{-4}$	18,5169	305,2803 - 1,1231
rs55728404	T/G	Olhos não verdes x olhos verdes	$9,9088 \times 10^{-4}$	18,5169	305,2803 - 1,1231
rs55728404	G	Olhos castanho-escuros x olhos azuis	$9,4060 \times 10^{-4}$	17,4959	288,7539 - 1,0601
rs55728404	T	Olhos azuis x olhos castanho-escuros	$9,4060 \times 10^{-4}$	17,4959	288,7539 - 1,0601
rs55728404	T	Olhos verdes x olhos não verdes	$1,3233 \times 10^{-3}$	17,3269	283,8936 - 1,0575
rs55728404	G	Olhos não verdes x olhos verdes	$1,3233 \times 10^{-3}$	17,3269	283,8936 - 1,0575
rs55728404	T/T	Pele clara x pele não clara	$4,2300 \times 10^{-7}$	10,9513	36,3200 - 3,3021
rs55728404	T/G	Pele não clara x pele clara	$4,2300 \times 10^{-7}$	10,9513	36,3200 - 3,3021
rs55728404	T	Pele clara x pele não clara	$7,5000 \times 10^{-7}$	10,0838	33,1223 - 3,0699
rs55728404	G	Pele não clara x pele clara	$7,5000 \times 10^{-7}$	10,0838	33,1223 - 3,0699
rs2675346 / rs2433354	T/T	Pele escura x pele clara	$2,2300 \times 10^{-8}$	94,3548	1623,2528 - 5,4846
rs2675346 / rs2433354	C/C	Cabelos loiro-escuros x cabelos pretos	$2,3000 \times 10^{-10}$	52,6154	399,7588 - 6,9251
rs2675346 / rs2433354	T	Cabelos pretos x cabelos loiro-escuros	$6,4000 \times 10^{-12}$	46,4655	341,5654 - 6,3210
rs2675346 / rs2433354	C	Cabelos loiro-escuros x cabelos pretos	$6,4000 \times 10^{-12}$	46,4655	341,5654 - 6,3210
rs2675346 / rs2433354	C/C	Cabelos loiro-claros x cabelos pretos	$1,9800 \times 10^{-5}$	42,7722	736,3253 - 2,4846
rs2675346 / rs2433354	T	Cabelos pretos x cabelos loiro-claros	$3,4800 \times 10^{-6}$	36,9142	613,1663 - 2,2223

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs2675346 / rs2433354	C	Cabelos loiro-claros x cabelos pretos	$3,4800 \times 10^{-6}$	36,9142	613,1663 - 2,2223
rs2675346 / rs2433354	T/T	Pele não clara x pele clara	$2,3500 \times 10^{-5}$	33,3856	558,1019 - 1,9971
rs2675346 / rs2433354	T/T	Cabelos pretos x cabelos castanho-claros	$6,9200 \times 10^{-5}$	31,2968	531,1506 - 1,8441
rs2675346 / rs2433354	T/T	Cabelos pretos x cabelos não pretos	$4,2700 \times 10^{-9}$	30,3377	134,7812 - 6,8286
rs2675346 / rs2433354	C/C	Olhos azuis x olhos castanho-escuros	$1,0500 \times 10^{-7}$	28,9796	215,7838 - 3,8919
rs2675346 / rs2433354	C/C	Pele clara x pele escura	$2,2900 \times 10^{-19}$	28,3273	64,0585 - 12,5266
rs2675346 / rs2433354	T	Olhos castanho-escuros x olhos azuis	$1,4900 \times 10^{-8}$	28,0757	204,7825 - 3,8492
rs2675346 / rs2433354	C	Olhos azuis x olhos castanho-escuros	$1,4900 \times 10^{-8}$	28,0757	204,7825 - 3,8492
rs2675346 / rs2433354	C/T	Cabelos pretos x cabelos loiro-escuros	$1,9500 \times 10^{-6}$	26,2545	199,5395 - 3,4545
rs2675346 / rs2433354	C/C	Cabelos castanho-claros x cabelos pretos	$1,6500 \times 10^{-13}$	23,8846	70,9604 - 8,0393
rs2675346 / rs2433354	C	Pele clara x pele escura	$2,3100 \times 10^{-24}$	23,3471	47,2028 - 11,5478
rs2675346 / rs2433354	T	Pele escura x pele clara	$2,3100 \times 10^{-24}$	23,3471	47,2028 - 11,5478
rs2675346 / rs2433354	C/T	Cabelos pretos x cabelos loiro-claros	$1,0089 \times 10^{-3}$	21,5045	370,2831 - 1,2489
rs2675346 / rs2433354	T	Cabelos pretos x cabelos castanho-claros	$3,8700 \times 10^{-16}$	21,4224	60,4283 - 7,5944
rs2675346 / rs2433354	C	Cabelos castanho-claros x cabelos pretos	$3,8700 \times 10^{-16}$	21,4224	60,4283 - 7,5944
rs2675346 / rs2433354	C/T	Olhos castanho-escuros x olhos azuis	$1,8700 \times 10^{-5}$	19,2982	144,0625 - 2,5851
rs2675346 / rs2433354	C	Olhos azuis x olhos não azuis	$6,5600 \times 10^{-5}$	14,9795	108,7838 - 2,0627
rs2675346 / rs2433354	T	Olhos não azuis x olhos azuis	$6,5600 \times 10^{-5}$	14,9795	108,7838 - 2,0627
rs2675346 / rs2433354	T/T	Cabelos pretos x cabelos castanho-escuros	$1,0600 \times 10^{-5}$	14,5455	64,9309 - 3,2584
rs2675346 / rs2433354	C	Cabelos loiro-escuros x cabelos não loiro-escuros	$1,0103 \times 10^{-4}$	14,1429	102,7709 - 1,9463
rs2675346 / rs2433354	T	Cabelos não loiro-escuros x cabelos loiro-escuros	$1,0103 \times 10^{-4}$	14,1429	102,7709 - 1,9463
rs2675346 / rs2433354	C/C	Olhos azuis x olhos não azuis	$2,8590 \times 10^{-4}$	14,0625	103,7835 - 1,9054
rs2675346 / rs2433354	T/T	Pele escura x pele não escura	$4,0500 \times 10^{-7}$	14,0435	39,2387 - 5,0261
rs2675346 / rs2433354	C/C	Cabelos ruivos x cabelos pretos	$2,6935 \times 10^{-3}$	13,8462	112,6731 - 1,7015

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs2675346 / rs2433354	C/C	Cabelos loiro-escuros x cabelos não loiro-escuros	$4,6924 \times 10^{-4}$	13,2558	97,9509 - 1,7939
rs2675346 / rs2433354	T	Cabelos pretos x cabelos ruivos	$1,4009 \times 10^{-3}$	12,6724	96,2865 - 1,6678
rs2675346 / rs2433354	C	Cabelos ruivos x cabelos pretos	$1,4009 \times 10^{-3}$	12,6724	96,2865 - 1,6678
rs2675346 / rs2433354	C/T	Pele escura x pele clara	$1,4900 \times 10^{-10}$	12,1136	26,9668 - 5,4415
rs2675346 / rs2433354	C/T	Cabelos pretos x cabelos castanho-claros	$6,8100 \times 10^{-8}$	11,9182	35,4296 - 4,0092
rs2675346 / rs2433354	T/T	Olhos castanho-escuros x olhos não castanho-escuros	$8,9600 \times 10^{-5}$	11,2941	49,8378 - 2,5594
rs2675346 / rs2433354	C/T	Olhos não azuis x olhos azuis	$2,4520 \times 10^{-3}$	10,5109	77,7606 - 1,4208
rs2459391	A/A	Olhos verdes x olhos castanho-escuros	$5,7800 \times 10^{-6}$	34,1136	565,1026 - 2,0593
rs2459391	A/G	Olhos castanho-escuros x olhos verdes	$5,7800 \times 10^{-6}$	34,1136	565,1026 - 2,0593
rs2459391	A/G	Cabelos pretos x cabelos loiro-escuros	$2,7700 \times 10^{-5}$	32,6692	550,5407 - 1,9386
rs2459391	A/A	Cabelos loiro-escuros x cabelos pretos	$2,7700 \times 10^{-5}$	32,6692	550,5407 - 1,9386
rs2459391	A	Olhos verdes x olhos castanho-escuros	$1,1200 \times 10^{-5}$	30,3748	499,3154 - 1,8478
rs2459391	G	Olhos castanho-escuros x olhos verdes	$1,1200 \times 10^{-5}$	30,3748	499,3154 - 1,8478
rs2459391	A/G	Cabelos pretos x cabelos castanho-claros	$4,2000 \times 10^{-7}$	29,4545	222,8572 - 3,8929
rs2459391	A/A	Cabelos castanho-claros x cabelos pretos	$4,2000 \times 10^{-7}$	29,4545	222,8572 - 3,8929
rs2459391	G	Cabelos pretos x cabelos loiro-escuros	$6,9700 \times 10^{-5}$	27,0690	449,5755 - 1,6298
rs2459391	A	Cabelos loiro-escuros x cabelos pretos	$6,9700 \times 10^{-5}$	27,0690	449,5755 - 1,6298
rs2459391	A/A	Pele clara x pele escura	$1,0200 \times 10^{-9}$	25,8000	91,8473 - 7,2472
rs2459391	A/G	Pele escura x pele clara	$1,0200 \times 10^{-9}$	25,8000	91,8473 - 7,2472
rs2459391	G	Cabelos pretos x cabelos castanho-claros	$1,1800 \times 10^{-6}$	24,6226	183,5056 - 3,3038
rs2459391	A	Cabelos castanho-claros x cabelos pretos	$1,1800 \times 10^{-6}$	24,6226	183,5056 - 3,3038
rs2459391	A	Pele clara x pele escura	$3,0900 \times 10^{-9}$	21,2449	73,6087 - 6,1317
rs2459391	G	Pele escura x pele clara	$3,0900 \times 10^{-9}$	21,2449	73,6087 - 6,1317
rs2459391	A/G	Olhos castanho-escuros x olhos azuis	$5,5936 \times 10^{-4}$	20,3700	339,6814 - 1,2215

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs2459391	A/A	Olhos azuis x olhos castanho-escuros	$5,5936 \times 10^{-4}$	20,3700	339,6814 - 1,2215
rs2459391	A/A	Olhos verdes x olhos não verdes	$5,6292 \times 10^{-4}$	19,0784	314,3920 - 1,1577
rs2459391	A/G	Olhos não verdes x olhos verdes	$5,6292 \times 10^{-4}$	19,0784	314,3920 - 1,1577
rs2459391	G	Olhos castanho-escuros x olhos azuis	$9,1773 \times 10^{-4}$	18,0933	298,4273 - 1,0970
rs2459391	A	Olhos azuis x olhos castanho-escuros	$9,1773 \times 10^{-4}$	18,0933	298,4273 - 1,0970
rs2459391	A	Olhos verdes x olhos não verdes	$7,5782 \times 10^{-4}$	17,8187	291,8127 - 1,0880
rs2459391	G	Olhos não verdes x olhos verdes	$7,5782 \times 10^{-4}$	17,8187	291,8127 - 1,0880
rs2459391	A/A	Pele clara x pele não clara	$2,2800 \times 10^{-7}$	11,3371	37,5508 - 3,4228
rs2459391	A/G	Pele não clara x pele clara	$2,2800 \times 10^{-7}$	11,3371	37,5508 - 3,4228
rs2459391	A/G	Cabelos pretos x cabelos não pretos	$1,0000 \times 10^{-10}$	10,5248	22,2925 - 4,9690
rs2459391	A/A	Cabelos não pretos x cabelos pretos	$1,0000 \times 10^{-10}$	10,5248	22,2925 - 4,9690
rs2459391	A	Pele clara x pele não clara	$4,2000 \times 10^{-7}$	10,4070	34,1381 - 3,1726
rs2459391	G	Pele não clara x pele clara	$4,2000 \times 10^{-7}$	10,4070	34,1381 - 3,1726
rs2459391	A/G	Olhos castanho-escuros x olhos não castanho-escuros	$1,1800 \times 10^{-8}$	10,3368	27,1284 - 3,9386
rs2459391	A/A	Olhos não castanho-escuros x olhos castanho-escuros	$1,1800 \times 10^{-8}$	10,3368	27,1284 - 3,9386
rs79856063	T/C	Cabelos pretos x cabelos não pretos	$1,3538 \times 10^{-2}$	22,7790	445,1338 - 1,1657
rs79856063	T/T	Cabelos não pretos x cabelos pretos	$1,3538 \times 10^{-2}$	22,7790	445,1338 - 1,1657
rs79856063	C	Cabelos pretos x cabelos não pretos	$1,3708 \times 10^{-2}$	22,4496	436,6315 - 1,1543
rs79856063	T	Cabelos não pretos x cabelos pretos	$1,3708 \times 10^{-2}$	22,4496	436,6315 - 1,1543
rs2433355	T/T	Cabelos pretos x cabelos loiro-claros	$1,8172 \times 10^{-2}$	34,6296	760,6935 - 1,5765
rs2433355	C/T	Cabelos loiro-claros x cabelos pretos	$1,8172 \times 10^{-2}$	34,6296	760,6935 - 1,5765
rs2433355	T	Cabelos pretos x cabelos loiro-claros	$1,8734 \times 10^{-2}$	32,7193	699,1105 - 1,5313
rs2433355	C	Cabelos loiro-claros x cabelos pretos	$1,8734 \times 10^{-2}$	32,7193	699,1105 - 1,5313
rs2433355	T/T	Cabelos castanho-claros x cabelos loiro-claros	$2,7429 \times 10^{-2}$	27,2222	599,0609 - 1,2370

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs2433355	C/T	Cabelos loiro-claros x cabelos castanho-claros	$2,7429 \times 10^{-2}$	27,2222	599,0609 - 1,2370
rs2433355	T	Cabelos castanho-claros x cabelos loiro-claros	$2,8247 \times 10^{-2}$	25,7018	549,6717 - 1,2018
rs2433355	C	Cabelos loiro-claros x cabelos castanho-claros	$2,8247 \times 10^{-2}$	25,7018	549,6717 - 1,2018
rs2433355	C/T	Olhos azuis x olhos não azuis	$3,0271 \times 10^{-3}$	12,3604	57,3619 - 2,6634
rs2433355	T/T	Olhos não azuis x olhos azuis	$3,0271 \times 10^{-3}$	12,3604	57,3619 - 2,6634
rs2433355	C	Olhos azuis x olhos não azuis	$3,2040 \times 10^{-3}$	11,7778	53,5900 - 2,5885
rs2433355	T	Olhos não azuis x olhos azuis	$3,2040 \times 10^{-3}$	11,7778	53,5900 - 2,5885
rs2433355	C/T	Cabelos loiro-claros x cabelos não loiro-claros	$2,6351 \times 10^{-2}$	11,2923	63,7344 - 2,0007
rs2433355	T/T	Cabelos não loiro-claros x cabelos loiro-claros	$2,6351 \times 10^{-2}$	11,2923	63,7344 - 2,0007
rs2433355	T/T	Cabelos castanho-escuros x cabelos loiro-claros	$4,5860 \times 10^{-2}$	10,7692	82,8787 - 1,3994
rs2433355	C/T	Cabelos loiro-claros x cabelos castanho-escuros	$4,5860 \times 10^{-2}$	10,7692	82,8787 - 1,3994
rs2433355	C	Cabelos loiro-claros x cabelos não loiro-claros	$2,7033 \times 10^{-2}$	10,5571	56,7970 - 1,9623
rs2433355	T	Cabelos não loiro-claros x cabelos loiro-claros	$2,7033 \times 10^{-2}$	10,5571	56,7970 - 1,9623
rs2433355	T	Cabelos castanho-escuros x cabelos loiro-claros	$4,6964 \times 10^{-2}$	10,0714	74,2747 - 1,3657
rs2433355	C	Cabelos loiro-claros x cabelos castanho-escuros	$4,6964 \times 10^{-2}$	10,0714	74,2747 - 1,3657
rs79338096	A/A	Olhos verdes x olhos castanho-escuros	$2,2289 \times 10^{-3}$	16,9736	285,7134 - 1,0084
rs79338096	A/C	Olhos castanho-escuros x olhos verdes	$2,2289 \times 10^{-3}$	16,9736	285,7134 - 1,0084
rs79338096	A/C	Cabelos pretos x cabelos castanho-claros	$3,6994 \times 10^{-3}$	11,7000	91,6804 - 1,4931
rs79338096	A/A	Cabelos castanho-claros x cabelos pretos	$3,6994 \times 10^{-3}$	11,7000	91,6804 - 1,4931
rs79338096	C	Cabelos pretos x cabelos castanho-claros	$4,4526 \times 10^{-3}$	10,8960	84,2913 - 1,4085
rs79338096	A	Cabelos castanho-claros x cabelos pretos	$4,4526 \times 10^{-3}$	10,8960	84,2913 - 1,4085
rs79338096	A/A	Pele clara x pele escura	$2,5416 \times 10^{-4}$	10,5306	40,4037 - 2,7446
rs79338096	A/C	Pele escura x pele clara	$2,5416 \times 10^{-4}$	10,5306	40,4037 - 2,7446
rs189640907	A/T	Cabelos castanho-claros x cabelos não castanho-claros	$2,2406 \times 10^{-2}$	13,4143	130,8798 - 1,3749

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs189640907	A/A	Cabelos não castanho-claros x cabelos castanho-claros	$2,2406 \times 10^{-2}$	13,4143	130,8798 - 1,3749
rs189640907	T	Cabelos castanho-claros x cabelos não castanho-claros	$2,2728 \times 10^{-2}$	13,1538	127,3784 - 1,3583
rs189640907	A	Cabelos não castanho-claros x cabelos castanho-claros	$2,2728 \times 10^{-2}$	13,1538	127,3784 - 1,3583
rs2433356	G/G	Pele clara x pele escura	$1,5100 \times 10^{-14}$	29,6129	82,7940 - 10,5916
rs2433356	A/A	Pele escura x pele clara	$3,5454 \times 10^{-3}$	28,9817	546,8477 - 1,5360
rs2433356	G	Pele clara x pele escura	$8,3700 \times 10^{-16}$	25,1647	66,6444 - 9,5021
rs2433356	A	Pele escura x pele clara	$8,3700 \times 10^{-16}$	25,1647	66,6444 - 9,5021
rs2433356	G/G	Olhos verdes x olhos castanho-escuros	$8,9500 \times 10^{-7}$	24,6774	182,9951 - 3,3278
rs2433356	G	Olhos verdes x olhos castanho-escuros	$1,8000 \times 10^{-7}$	24,3449	178,0116 - 3,3294
rs2433356	A	Olhos castanho-escuros x olhos verdes	$1,8000 \times 10^{-7}$	24,3449	178,0116 - 3,3294
rs2433356	G/G	Cabelos loiro-escuros x cabelos pretos	$9,2400 \times 10^{-6}$	22,9310	174,4973 - 3,0134
rs2433356	G/A	Pele escura x pele clara	$1,0100 \times 10^{-11}$	22,3429	62,7893 - 7,9504
rs2433356	A	Cabelos pretos x cabelos loiro-escuros	$3,8500 \times 10^{-6}$	21,7724	161,3418 - 2,9381
rs2433356	G	Cabelos loiro-escuros x cabelos pretos	$3,8500 \times 10^{-6}$	21,7724	161,3418 - 2,9381
rs2433356	G/G	Cabelos castanho-claros x cabelos pretos	$1,2100 \times 10^{-8}$	21,4224	92,8519 - 4,9425
rs2433356	G/A	Olhos castanho-escuros x olhos verdes	$6,5200 \times 10^{-6}$	20,4000	151,7181 - 2,7430
rs2433356	A	Cabelos pretos x cabelos castanho-claros	$1,4100 \times 10^{-9}$	20,3586	85,7484 - 4,8336
rs2433356	G	Cabelos castanho-claros x cabelos pretos	$1,4100 \times 10^{-9}$	20,3586	85,7484 - 4,8336
rs2433356	A/A	Cabelos pretos x cabelos não pretos	$9,5626 \times 10^{-4}$	20,2069	170,1209 - 2,4002
rs2433356	G/G	Cabelos loiro-claros x cabelos pretos	$2,2432 \times 10^{-3}$	18,8120	324,1986 - 1,0916
rs2433356	A	Cabelos pretos x cabelos loiro-claros	$1,8372 \times 10^{-3}$	17,3986	290,6245 - 1,0416
rs2433356	G	Cabelos loiro-claros x cabelos pretos	$1,8372 \times 10^{-3}$	17,3986	290,6245 - 1,0416
rs2433356	G/A	Cabelos pretos x cabelos loiro-escuros	$1,6697 \times 10^{-4}$	17,2188	131,5620 - 2,2536
rs2433356	G/A	Cabelos pretos x cabelos castanho-claros	$1,3100 \times 10^{-6}$	16,0859	70,1144 - 3,6905

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs2433356	G/G	Olhos azuis x olhos castanho-escuros	$2,5313 \times 10^{-4}$	14,5161	108,7091 - 1,9384
rs2433356	A	Olhos castanho-escuros x olhos azuis	$1,2697 \times 10^{-4}$	14,3937	105,7611 - 1,9589
rs2433356	G	Olhos azuis x olhos castanho-escuros	$1,2697 \times 10^{-4}$	14,3937	105,7611 - 1,9589
rs2433356	G	Olhos verdes x olhos não verdes	$8,5200 \times 10^{-5}$	14,2708	103,8772 - 1,9606
rs2433356	A	Olhos não verdes x olhos verdes	$8,5200 \times 10^{-5}$	14,2708	103,8772 - 1,9606
rs2433356	G/G	Olhos verdes x olhos não verdes	$2,1017 \times 10^{-4}$	13,6000	100,1129 - 1,8475
rs2433356	G/A	Olhos castanho-escuros x olhos azuis	$1,4438 \times 10^{-3}$	12,0000	90,1276 - 1,5977
rs2433356	G/A	Olhos não verdes x olhos verdes	$9,4906 \times 10^{-4}$	11,5000	84,8804 - 1,5581
rs2433356	G	Pele clara x pele não clara	$9,0400 \times 10^{-11}$	10,5000	26,5202 - 4,1572
rs2433356	A	Pele não clara x pele clara	$9,0400 \times 10^{-11}$	10,5000	26,5202 - 4,1572
rs2433356	G/G	Pele clara x pele não clara	$1,8200 \times 10^{-9}$	10,2209	26,2940 - 3,9730
rs2433356	A/A	Manchas expostas sim x manchas expostas não	$7,8765 \times 10^{-4}$	21,1429	2,5105 - 178,0610
rs138961421	A/C	Cabelos pretos x cabelos não pretos	$1,3538 \times 10^{-2}$	22,7790	445,1338 - 1,1657
rs138961421	A/A	Cabelos não pretos x cabelos pretos	$1,3538 \times 10^{-2}$	22,7790	445,1338 - 1,1657
rs138961421	C	Cabelos pretos x cabelos não pretos	$1,3708 \times 10^{-2}$	22,4496	436,6315 - 1,1543
rs138961421	A	Cabelos não pretos x cabelos pretos	$1,3708 \times 10^{-2}$	22,4496	436,6315 - 1,1543
rs1426654	G/G	Pele escura x pele clara	$2,7500 \times 10^{-22}$	214,1538	1634,7776 - 28,0539
rs1426654	A/A	Cabelos loiro-claros x cabelos pretos	$6,3600 \times 10^{-9}$	104,5349	1819,8590 - 6,0046
rs1426654	G/G	Cabelos pretos x cabelos castanho-claros	$4,9000 \times 10^{-11}$	89,2051	1485,1089 - 5,3582
rs1426654	G	Cabelos pretos x cabelos loiro-claros	$1,8800 \times 10^{-10}$	82,4843	1369,2825 - 4,9688
rs1426654	A	Cabelos loiro-claros x cabelos pretos	$1,8800 \times 10^{-10}$	82,4843	1369,2825 - 4,9688
rs1426654	A/A	Pele clara x pele escura	$1,5200 \times 10^{-29}$	75,5926	200,5499 - 28,4929
rs1426654	G/G	Olhos castanho-escuros x olhos verdes	$2,7800 \times 10^{-8}$	52,3360	862,4788 - 3,1758
rs1426654	G/G	Pele não clara x pele clara	$7,7000 \times 10^{-13}$	50,9268	373,2022 - 6,9494

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs1426654	G/G	Cabelos pretos x cabelos loiro-escuros	$5,7800 \times 10^{-7}$	47,9402	804,5198 - 2,8567
rs1426654	A	Pele clara x pele escura	$1,2200 \times 10^{-46}$	45,7303	84,6774 - 24,6968
rs1426654	G	Pele escura x pele clara	$1,2200 \times 10^{-46}$	45,7303	84,6774 - 24,6968
rs1426654	A/A	Cabelos loiro-escuros x cabelos pretos	$3,1100 \times 10^{-14}$	41,1429	147,1138 - 11,5063
rs1426654	A/A	Cabelos ruivos x cabelos pretos	$1,4900 \times 10^{-5}$	34,2857	283,4530 - 4,1471
rs1426654	G	Cabelos pretos x cabelos loiro-escuros	$1,6900 \times 10^{-18}$	33,8608	111,3093 - 10,3006
rs1426654	A	Cabelos loiro-escuros x cabelos pretos	$1,6900 \times 10^{-18}$	33,8608	111,3093 - 10,3006
rs1426654	G/G	Olhos castanho-escuros x olhos azuis	$1,3600 \times 10^{-5}$	31,2510	518,4397 - 1,8838
rs1426654	G	Cabelos pretos x cabelos ruivos	$9,8800 \times 10^{-7}$	28,4430	215,9288 - 3,7466
rs1426654	A	Cabelos ruivos x cabelos pretos	$9,8800 \times 10^{-7}$	28,4430	215,9288 - 3,7466
rs1426654	G/G	Cabelos não castanho-claros x cabelos castanho-claros	$2,2000 \times 10^{-5}$	27,4068	449,6699 - 1,6704
rs1426654	G/G	Olhos castanho-escuros x olhos não castanho-escuros	$2,6100 \times 10^{-15}$	26,8022	87,9897 - 8,1641
rs1426654	G/G	Olhos não verdes x olhos verdes	$7,1100 \times 10^{-5}$	25,5306	419,0963 - 1,5553
rs1426654	G/G	Pele escura x pele não escura	$6,1300 \times 10^{-19}$	22,5882	46,0081 - 11,0900
rs1426654	A	Cabelos loiro-claros x cabelos não loiro-claros	$1,6711 \times 10^{-4}$	22,0046	361,5559 - 1,3392
rs1426654	G	Cabelos não loiro-claros x cabelos loiro-claros	$1,6711 \times 10^{-4}$	22,0046	361,5559 - 1,3392
rs1426654	A/A	Cabelos castanho-claros x cabelos pretos	$4,5300 \times 10^{-17}$	21,6000	49,3118 - 9,4614
rs1426654	A/A	Cabelos loiro-claros x cabelos castanho-escuros	$1,1165 \times 10^{-3}$	21,4615	365,7988 - 1,2592
rs1426654	A/A	Pele não escura x pele escura	$1,4100 \times 10^{-18}$	21,0347	50,6037 - 8,7436
rs1426654	A/G	Cabelos pretos x cabelos loiro-claros	$2,1248 \times 10^{-3}$	20,5752	354,3713 - 1,1946
rs1426654	A/A	Cabelos loiro-claros x cabelos não loiro-claros	$7,3312 \times 10^{-4}$	20,5056	345,3217 - 1,2176
rs1426654	G/G	Pele intermediária x pele clara	$4,9100 \times 10^{-5}$	20,1739	154,0059 - 2,6427
rs1426654	G	Cabelos castanho-escuros x cabelos loiro-claros	$7,3960 \times 10^{-4}$	20,0490	332,1331 - 1,2102
rs1426654	A	Cabelos loiro-claros x cabelos castanho-escuros	$7,3960 \times 10^{-4}$	20,0490	332,1331 - 1,2102

<i>Sítio de variação</i>	<i>Alelo / Genótipo</i>	<i>Característica</i>	<i>Probabilidade do teste exato de Fisher</i>	<i>Odds Ratio</i>	<i>IC 95%</i>
rs1426654	G/G	Cabelos pretos x cabelos loiro-claros	$2,2432 \times 10^{-3}$	18,8120	324,1986 - 1,0916
rs1426654	G	Cabelos pretos x cabelos castanho-claros	$4,8500 \times 10^{-24}$	18,4203	37,2770 - 9,1023
rs1426654	A	Cabelos castanho-claros x cabelos pretos	$4,8500 \times 10^{-24}$	18,4203	37,2770 - 9,1023
rs1426654	G	Olhos castanho-escuros x olhos azuis	$2,3700 \times 10^{-13}$	15,5585	43,4695 - 5,5687
rs1426654	A	Olhos azuis x olhos castanho-escuros	$2,3700 \times 10^{-13}$	15,5585	43,4695 - 5,5687
rs1426654	A/A	Olhos azuis x olhos castanho-escuros	$1,1700 \times 10^{-9}$	14,8000	43,4567 - 5,0404
rs1426654	A/A	Pele clara x pele não clara	$1,7000 \times 10^{-26}$	13,8279	24,2228 - 7,8938
rs1426654	G	Pele escura x pele não escura	$1,2700 \times 10^{-31}$	12,6604	19,9505 - 8,0342
rs1426654	A	Pele não escura x pele escura	$1,2700 \times 10^{-31}$	12,6604	19,9505 - 8,0342
rs1426654	A	Pele clara x pele não clara	$5,9300 \times 10^{-35}$	12,6055	20,8034 - 7,6381
rs1426654	G	Pele não clara x pele clara	$5,9300 \times 10^{-35}$	12,6055	20,8034 - 7,6381
rs1426654	G/G	Cabelos pretos x cabelos não pretos	$3,6200 \times 10^{-14}$	12,0690	23,8524 - 6,1067
rs1426654	G/G	Pele escura x pele intermediária	$4,4800 \times 10^{-11}$	10,6154	22,0703 - 5,1058
rs1426654	G/G	Olhos castanho-escuros x olhos castanho-claros	$6,4800 \times 10^{-7}$	10,5962	35,1868 - 3,1910

Apêndice B – Frequências globais dos alelos referência e alternativo dos sítios de variação rs553733344, rs55728404, rs2675346 e rs1426654.

Apêndice B – Frequências globais dos alelos referência e alternativo dos sítios de variação que apresentaram associações com fenótipos na amostra populacional de Ribeirão Preto explicadas pela interação com Fatores de Transcrição ou alteração na sequência primária da proteína NCKX5.

Sítio de variação	REF	ALT	RBP		AFR		AMR		EAS		EUR		SAS	
			REF	ALT	REF	ALT	REF	ALT	REF	ALT	REF	ALT	REF	ALT
rs553733344	C	CA	0,9517	0,0483	0,7490	0,2510	0,8482	0,1518	0,9364	0,0636	0,9990	0,0010	0,9334	0,0666
rs55728404	T	G	0,9529	0,0471	0,7470	0,2530	0,8482	0,1518	0,9364	0,0636	0,9990	0,0010	0,9334	0,0666
rs2675346	C	T	0,8588	0,1412	0,3075	0,6925	0,6825	0,3175	0,5080	0,4920	0,9990	0,0010	0,7859	0,2141
rs1426654	A	G	0,7443	0,2557	0,0546	0,9454	0,4484	0,5516	0,0119	0,9881	0,9970	0,0030	0,6855	0,3145

Anexos

Anexo A – Termo de Consentimento Livre e Esclarecido**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 custos. 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 nº _____,
residente na _____ nº
_____, 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 – 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	117
2. QUANTIFICAÇÃO E VERIFICAÇÃO DA QUALIDADE DO DNA	117
3. PREPARO DOS REAGENTES PARA ENRIQUECIMENTO DE ALVO	118
4. ENRIQUECIMENTO DO ALVO E PREPARO DA BIBLIOTECA	118
4.1. Digestão e hibridação do DNA	119
4.2. Captura, ligação, eluição e PCR	120
4.3. Purificação da biblioteca	121
5. CONTROLE DE QUALIDADE E QUANTIFICAÇÃO DA BIBLIOTECA	121
6. SEQUENCIAMENTO	122
7. REFERÊNCIAS	123

1. EXTRAÇÃO DE DNA GENÔMICO

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

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

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

2. QUANTIFICAÇÃO E VERIFICAÇÃO DA QUALIDADE DO DNA

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

2,0. Além disso, é realizada a eletroforese em gel de agarose, a fim de verificar o grau de degradação do material, considerando como indicativo de degradação qualquer banda abaixo de 2,5kb (*Agilent Technologies*).

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

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

3. PREPARO DOS REAGENTES PARA ENRIQUECIMENTO DE ALVO

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

4. ENRIQUECIMENTO DO ALVO E PREPARO DA BIBLIOTECA

Após a quantificação, cada amostra de DNA é processada individualmente, gerando uma biblioteca contendo todos os amplicons. A captura das regiões-alvo utiliza reagentes do kit de enriquecimento de alvo customizado *HaloPlex* (*Agilent Technologies*) e segue o

protocolo sugerido pelo fabricante. Resumidamente, o protocolo consiste dos seguintes passos: 1) digestão do DNA genômico com enzimas de restrição; 2) hibridação dos fragmentos de DNA a sondas biotiniladas cujos finais são complementares a esses fragmentos-alvo (durante esse passo os fragmentos são circularizados e índices, que irão identificar cada indivíduo durante o sequenciamento, são adicionados); 3) captura dos fragmentos-alvo utilizando *beads* magnéticas recobertas por estreptavidina; 4) ligação dos fragmentos circularizados; 5) amplificação por PCR dos fragmentos-alvo capturados para montagem da biblioteca.

4.1. Digestão e hibridação do DNA

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

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

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

O DNA digerido de cada amostra (80µL somando todos os 8 tubos de digestão) é então transferido para os microtubos contendo o mix de hibridação (70µL) e seus índices correspondentes (10µL). Essas soluções são incubadas a 54°C durante 16 horas. O tempo de

duração da hibridação é determinado de acordo com a densidade de sondas do ensaio. Segundo instruções do fabricante (*Haloplex, Agilent Technologies*), ensaios que atingem cobertura de entre 500kb - 5Mb exigem 16 horas de hibridação, enquanto que ensaios com cobertura de entre 1-500kb exigem apenas 3 horas de hibridação. Apesar de o presente ensaio estar voltado para a cobertura de apenas 488,7kb, os ensaios iniciais demonstraram a necessidade de um tempo de hibridação de 16 horas.

4.2. Captura, ligação, eluição e PCR

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

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

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

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

seguida por 20 ciclos de 98°C por 30 segundos, 60°C por 30 segundos e 72°C por 1 minuto, e um passo final a 72°C por 10 minutos. Após o término da reação de PCR o produto é mantido a 8°C.

4.3. Purificação da biblioteca

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

5. CONTROLE DE QUALIDADE E QUANTIFICAÇÃO DA BIBLIOTECA

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

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

Figura 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 C – 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	125
1.1. Análise primária	125
1.2. Análise secundária	126
1.2.1. Avaliação da qualidade	126
1.2.2. Trimagem	126
1.2.3. Alinhamento ao genoma de referência	128
1.2.3.1. SAM	130
1.2.3.2. BAM	131
1.3. Análise terciária	132
1.3.1. Determinação de variantes	132
1.3.1.1. VCF	136
1.3.2. Remoção de genótipos de baixa qualidade	138
2. REFERÊNCIAS	140

1. ANÁLISE DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

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

1.1. Análise primária

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

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

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

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

1.2. Análise secundária

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

1.2.1. Avaliação da qualidade

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

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

1.2.2. Trimagem

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

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

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

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

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

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


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

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

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

1.2.3. Alinhamento ao genoma de referência

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

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

Antes de realizar o alinhamento, porém, é preciso indexar o genoma de referência por meio do índice FM (Ferragina, Manzini, Veli, & Navarro, 2004). Neste caso, utilizamos um genoma já indexado (*Homo sapiens*, UCSC, hg19), disponibilizado pela *Illumina* através do *iGenomes* em: https://support.illumina.com/sequencing/sequencing_software/igenome.html.

Pode-se também indexar localmente o genoma de interesse por meio de ferramentas específicas. Para tanto, é necessário gerar três diferentes arquivos, arquivo índice, arquivo fasta índice e um dicionário de sequência, os quais podem ser obtidos por meio dos programas BWA, Samtools e Picard, respectivamente.

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

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

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

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

	valor SM será tratado como pertencendo a uma mesma amostra.
--	---

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

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

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

1.2.3.1. SAM

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

Arquivos SAM possuem uma seção com o cabeçalho (que é opcional) e uma seção de alinhamento. As linhas no cabeçalho iniciam com “@” seguido por um código de duas letras.

Por sua vez, cada *read* alinhada à referência é representada em uma linha da seção de alinhamento.

Podem estar presentes ainda linhas opcionais com marcações predefinidas (*The SAM/BAM Format Specification Working Group, Sequence Alignment/Map Format Specification*, 2015, <http://samtools.github.io/hts-specs/SAMv1.pdf>). Algumas destas marcações são geradas pelo BWA (aquelas começando com “X”).

1.2.3.2. BAM

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

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

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

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

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

```
PIG0001.sorted.bam
```

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

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

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

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

```
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 \
  --emitRefConfidence GVCF \
  --dbsnp dbSNP.vcf \
  -L targets.interval_list \
  -o output.raw.snps.indels.g.vcf \
  --dontUseSoftClippedBases \
  -drf DuplicateRead
```

UnifiedGenotyper:

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

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

Comandos utilizados no GATK	Função do comando executado
-R	indica o arquivo com o genoma de referência
-T	indica qual ferramenta do GATK será utilizada, isto é, <i>HaplotypeCaller</i> ou <i>UnifiedGenotyper</i>

-I	indica arquivo de texto com lista de amostras (formato BAM) a ser utilizado pelo GATK
--dbsnp	indica arquivo (dbSNP.vcf) contendo a identificação dos rs de cada ponto de variação oficialmente descrito, a ser utilizada para preenchimento da coluna ID do <i>output</i>
-stand_call_conf	é o limiar mínimo de confiança (na escala Phred) utilizado para separar determinações de genótipo realizadas com maior e menor confiança. Apenas determinações de genótipo emitidas com confiança $\geq$ limiar mínimo são consideradas. O valor default recomendado é de 30

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

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

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

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


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

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

1.3.1.1. VCF

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

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

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

Além do cabeçalho, um arquivo VCF contém as linhas com os dados das amostras, onde estão contidas em cada linha informações sobre uma única variante. As linhas com os

dados são listadas seguindo o designado no cabeçalho do arquivo VCF. As oito primeiras colunas do arquivo (até INFO) representam as propriedades observadas dos sítios variantes (ou invariantes). Informações específicas à amostra são apresentadas na coluna FORMAT e nas seguintes.

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

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

Informações presentes no corpo do arquivo	Descrição das abreviações apresentadas
CHROM e POS	contém as coordenadas genômicas nas quais as variantes ocorrem
ID	campo opcional, cujo conteúdo se baseia nas informações contidas em um banco de dados de referência como o dbSNP
REF e ALT	identificam o alelo referência e alternativo(s) observado(s) na(s) amostra(s). No caso de inserções, o alelo ALT fornece a sequência inserida, bem como a base anterior à inserção. Para deleções, o alelo ALT é representado pela base anterior à deleção
QUAL	é a probabilidade de que um polimorfismo (REF/ALT) exista de fato neste local indicado. No entanto, por ser este valor dependente da quantidade de dados analisada ele não representa uma medida útil para avaliação da qualidade de determinação de uma variante
FILTER	esse campo contém os filtros aplicados aos dados analisados. Se a variante passou pelos filtros aplicados o valor indicado é PASS. Caso nenhum filtro tenha sido aplicado será indicado “.”
INFO	informações descrevendo cada variante
DB	associado ao dbSNP
H3	associado ao HapMap3
VALIDATED	validado por experimentos contínuos
NA	número total de alelos nos genótipos identificados
AC	contagem alélica dentre os genótipos do alelo ALT, na mesma ordem em que listados
SVTYPE	tipo de variante estrutural (DEL para deleção, DUP para duplicação, INV para inversão)
END	posição final ocupada pela variante
IMPRECISE	indica que a posição ocupada pela variante

	não é precisa
CIPOS/CIEND	intervalo de confiança em torno da posição e posição final para variantes localizadas com baixa precisão
FORMAT	fornece a ordem das informações contidas nas colunas seguintes (amostras):
GT	indica o genótipo da amostra em cada sítio. Em diploides, são indicados os dois alelos de cada indivíduo, sendo “0” indicativo do alelo referência e “1” do alelo alternativo. O separador (/) indica que os genótipos não estão faseados e () indica genótipos em fase
AD	representa o número de <i>reads</i> que dá suporte a cada um dos alelos (alternativo e/ou referência) reportados
DP	indica o número total de <i>reads</i> que representam o genótipo reportado
GQ	indica qual a confiança (escala Phred) da determinação de genótipo realizada. Valores muito baixos são indicativos de pouca confiança no genótipo identificado
PL	é a probabilidade normalizada dos genótipos possíveis ou a probabilidade de o genótipo não estar correto. Assim, quanto menor o valor apresentado aqui, melhor
HP	identificadores de haplótipos eventualmente determinados pelo <i>ReadBackedPhasing</i>
PQ	qualidade de faseamento de haplótipos eventualmente determinados pelo <i>ReadBackedPhasing</i>

1.3.2. Remoção de genótipos de baixa qualidade

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

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

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

- Para genótipos homozigotos inferidos em segmentos cobertos por oito ou menos *reads*, é introduzido um alelo interrogado (*vcfx alpha* = 8). A definição final desse tipo de genótipo (homozigoto ou heterozigoto) será inferida posteriormente.
- Genótipos heterozigotos em que um dos alelos está altamente subrepresentado (proporção de *reads* igual ou menor que 1%), são considerados homozigotos para o alelo mais representado (*vcfx beta* = 0.01). Esse procedimento minimiza a influência de *reads* mapeadas incorretamente e da alta taxa de erros de sequenciamento geralmente introduzidos em dados de NGS. Essa correção só pode ser aplicada em situações de alta cobertura (100 ou mais *reads*).
- Para genótipos heterozigotos em que um dos alelos está relativamente subrepresentado (proporção de *reads* entre 1% e 20%), um alelo interrogado é introduzido representando o alelo subrepresentado (*vcfx delta* = 0.20). A definição final desse tipo de genótipo (homozigoto ou heterozigoto) será inferida posteriormente.
- Para genótipos heterozigotos inferidos com no máximo cinco *reads* em que um dos alelos está subrepresentado (Proporção de *reads* ente 20% e 40%), um alelo interrogado é inserido representando o alelo subrepresentado. No entanto, quando os dois alelos têm a mesma representatividade, ambos são considerados interrogados (*vcfx gamma* = 0.50). A definição final desse tipo de genótipo (homozigoto ou heterozigoto) será inferida posteriormente.

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

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

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

as regras empregadas pelo VCFx são novamente aplicadas, agora manualmente para determinar se essas variantes continuariam interrogadas ou se seriam resolvidas.

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

Após a análise pelo VCFx, se apenas a ferramenta *HaplotypeCaller* com o parâmetro -emitRefConfidence GVCF foi utilizada, os pontos de variação interrogados das amostras devem ser visualizados no programa IGV. É preciso analisar os arquivos BAMs destas amostras para avaliar a qualidade e contagem de *reads* alternativas que cobrem determinada posição (interrogada pelo programa VCFx). Isto porque o *HaplotypeCaller* aplica filtros extremamente rigorosos, descartando um número considerável de *reads* para suas análises. Portanto, analisar ponto a ponto os arquivos BAM muitas vezes pode auxiliar na compreensão do porquê as variantes foram interrogadas pelo programa VCFx.

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

2. REFERÊNCIAS

- Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Retrieved from <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Castelli, E. C., Mendes-Junior, C. T., Sabbagh, A., Porto, I. O. P., Garcia, A., Ramalho, J., ... Donadi, E. A. (2015). HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples. *Human Immunology*, 76(12), 945–953. <https://doi.org/10.1016/j.humimm.2015.06.016>
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), 491–8. <https://doi.org/10.1038/ng.806>
- Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research*, 8(3), 175–185. <https://doi.org/10.1101/gr.8.3.175>
- Ferragina, P., Manzini, G., Veli, M., & Navarro, G. (2004). An Alphabet-Friendly FM-Index, 150–160.

- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), 10–12. <https://doi.org/10.14806/ej.17.1.200>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. <https://doi.org/10.1038/nbt.1754>
- Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192. <https://doi.org/10.1093/bib/bbs017>
- Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., ... DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics* (Vol. 11, p. 11.10.1-11.10.33). Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/0471250953.bi1110s43>

Anexo D – Protocolo de análise estatística de dados de sequenciamento de nova geração

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

Protocolo de análise estatística 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 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. FASEAMENTO EXPERIMENTAL	143
2. FASEAMENTO COMPUTACIONAL	144
2.1. Faseamento completo	145
2.2. Faseamento fragmentado	146
3. DESEQUILÍBRIO DE LIGAÇÃO	147
4. FREQUÊNCIAS ALÉLICAS, EQUILÍBRIO DE HARDY-WEINBERG E HETEROZIGOSE ESPERADA	148
5. ASSOCIAÇÕES GENÓTIPO-FENÓTIPO	149
6. REFERÊNCIAS	149

1. FASEAMENTO EXPERIMENTAL

É utilizada a ferramenta *ReadBackedPhasing*, também incorporada no programa GATK (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013), a qual se utiliza da informação da leitura das *reads paired-end* para determinar a fase de alelos de variantes presentes em um mesmo par de *reads*.

A linha de comando a seguir é utilizada para realizar esse pré-faseamento (Tabela 1):

```
java -jar GenomeAnalysisTK.jar \
  -T ReadBackedPhasing \
  -R reference.fasta \
  -I reads.bam \
  --variant SNPs.vcf \
  -L SNPs.vcf \
  -o phased_SNPs.vcf \
  --phaseQualityThresh 500.0
```

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

Comandos utilizados no <i>ReadBackedPhasing</i>	Função do comando executado
-T	indica qual ferramenta do GATK será utilizada. Neste caso, o <i>ReadBackedPhasing</i>
-R	indica o arquivo com o genoma de referência
-I	indica arquivo de texto com lista de amostras (formato BAM) a ser utilizado pelo <i>ReadBackedPhasing</i>
--variant	arquivo VCF consenso (*)
-L	direciona o GATK a restringir o processamento dos dados às variantes especificadas no arquivo “ <i>SNPs.vcf</i> ”
-o	indica o <i>output</i> , que é um arquivo VCF processado contendo os sítios de variação originais, com fase determinada para alelos de sítios próximos
--phaseQualityThresh	valor que indica a qualidade mínima de faseamento a ser empregada pelo algoritmo para aceitar uma fase como verdadeira. O valor <i>default</i> recomendado é 30. O valor de 500 aqui utilizado assegura que apenas fases bem definidas pelo GATK sejam aceitas como verdadeira

(*) Arquivo gerado a partir da comparação dos arquivos gerados pelo *HaplotypeCaller* e *UnifiedGenotyper*, posteriormente processados pelo VCFx.0.10.1 (Castelli et al., 2015). Vide Protocolo de processamento computacional de dados de sequenciamento de nova geração.

No entanto, a informação obtida através dos dados de sequenciamento ainda não é suficiente para determinar a fase completa dos haplótipos. Por esse motivo, o resultado de faseamento obtido através do *ReadBackedPhasing* deve ser incorporado ao método de faseamento computacional PHASE 2.1 (M Stephens, Smith, & Donnelly, 2001; Matthew Stephens & Donnelly, 2003).

2. FASEAMENTO COMPUTACIONAL

A partir das diversas configurações de haplótipo possíveis para os genótipos identificados são estimadas as probabilidades de uma determinada configuração haplotípica. Dentre as abordagens estatísticas disponíveis para o faseamento de indivíduos não aparentados, optamos pela utilização do programa PHASE 2.1 (M Stephens et al., 2001; Matthew Stephens & Donnelly, 2003).

O PHASE implementa um método de estatística Bayesiana para reconstruir os haplótipos a partir dos dados de genotipagem da população (Matthew Stephens & Donnelly, 2003). O método empregado pelo programa se baseia na utilização de modelos coalescentes, segundo os quais os novos haplótipos seriam derivados de haplótipos antigos pelos processos de mutação e recombinação (o que ajuda a entender a grande semelhança observada entre os haplótipos) e no modelo de *Hidden Markov* (HMM), um processo probabilístico segundo o qual a distribuição dos estados futuros depende apenas do estado atual e não de estados prévios. São utilizados os seguintes parâmetros para as análises realizadas no PHASE: número de interações (1000), *thinning interval* (1), *burn-in value* (100), -X10 (o software faz uma última corrida utilizando 10x o número de interações especificado), -x10 (o software aplica o algoritmo 10 vezes e devolve o melhor resultado). Como muitos sítios de variação podem se mostrar não-polimórficos em uma dada amostra populacional, levando o PHASE a inferir haplótipos com valores muito baixos de probabilidade, os *singletons* (contagem alélica = 1) são excluídos desta análise.

A determinação da fase dos haplótipos pode ser feita por meio de duas abordagens: faseamento completo e/ou faseamento fragmentado. O faseamento completo geralmente é mais indicado para regiões menores, em que as informações obtidas pelo *ReadBackedPhasing*

são melhor aproveitadas pelo PHASE devido à (a) ausência de lacunas entre as regiões cobertas pelo ensaio de sequenciamento e (b) menor distância entre as variantes encontradas na região. Já o faseamento fragmentado é mais indicado para regiões maiores, principalmente quando há (a) lacunas no sequenciamento (íntrons, por exemplo) e (b) as variantes estão distantes umas das outras (diferentes éxons, por exemplo), impossibilitando o faseamento experimental pelo *ReadBackedPhasing* e tornando o faseamento computacional mais complexo. Nesse caso, o faseamento pode ser feito para diferentes regiões (por exemplo, promotora, 5'UTR, CDS, 3'UTR) separadamente e posteriormente é feita uma etapa final de faseamento unindo os haplótipos encontrados.

2.1. Faseamento completo

Uma primeira corrida é realizada utilizando o arquivo VCF final obtido à partir da comparação entre os arquivos gerados pelos algoritmos *HaplotypeCaller* e *UnifiedGenotyper*. Adicionalmente, foram desenvolvidos *scripts* (Castelli E.C.) para implementar uma metodologia que faz uso das informações de fase definidas pelo algoritmo *ReadBackedPhasing*, utilizando a opção *-k* do *software* PHASE. Nesta metodologia, para cada amostra são criados blocos de haplótipos de fase experimentalmente determinada. Em seguida, o programa PHASE realiza análises de inferência de haplótipos fixando cada um desses blocos e inferindo o restante do haplótipo de cada amostra, sendo o número de análises equivalente ao número máximo de blocos observado para as amostras que compõem o arquivo *input*. Além disso, o programa ainda faz uma última análise de inferências desconsiderando quaisquer informações do faseamento experimental. Com base nos resultados de haplótipos obtidos é realizada a maturação visual dos resultados, sendo selecionados indivíduos que tiveram o mesmo par de haplótipos inferido com probabilidade de pelo menos 80% em todas as corridas do programa.

Feita essa seleção, é realizada uma segunda corrida no PHASE especificando a fase completa dos indivíduos que tiveram seus haplótipos inferidos na corrida anterior (probabilidade > 80%), permitindo com isso que o programa melhorasse seu desempenho na inferência dos haplótipos dos indivíduos restantes. Para tanto, o arquivo contendo as fases conhecidas (.known) é editado, onde as informações de fase daqueles indivíduos selecionados anteriormente são substituídas por "0", o que indica que a fase para todos os *loci* de

determinado indivíduo é a mesma que a dos genótipos apresentados no arquivo *input* (.inp), o qual também é editado para conter a fase correta dos haplótipos desses indivíduos.

Após a determinação dos haplótipos, o arquivo out.pairs editado (contendo somente o par de haplótipos inferido com maior probabilidade para cada indivíduo) é utilizado como *input* para rodar o *script* phase_to_vcf.pl (Castelli E.C.), que gera como *output* um arquivo VCF faseado. Neste momento, as linhas com os *singletons* que tiveram a fase informada pelo *ReadBackedPhasing* podem ser reinseridas no arquivo VCF faseado, caso o pesquisador esteja certo de que aquela região em que o *singleton* se encontra tenha sido adequadamente sequenciada em todas as amostras.

2.2. Faseamento fragmentado

Primeiramente é preciso decidir em quantos fragmentos a região cromossômica será dividida. Essa decisão deve ser tomada levando em consideração a quantidade de variantes presentes em cada região e a existência de lacunas entre regiões cobertas (por exemplo, promotora, 5'UTR, éxons, 3'UTR).

Após essa separação, cada região deve ser faseada utilizando o software PHASE com os parâmetros definidos anteriormente. Os *scripts* citados no faseamento completo desenvolvidos por Castelli E.C. podem ser implementados nesse momento caso, para essa região, o *ReadBackedPhasing* tenha sido informativo. Terminado o faseamento, são aceitos como corretos os haplótipos inferidos com probabilidade superior a 80%, sendo os pares de haplótipos dos demais indivíduos considerados como *missing regions*.

Após ter determinado os haplótipos para cada fragmento, para definir o haplótipo estendido (completo) é preciso construir um novo arquivo de *input* para o PHASE em que cada fragmento faseado anteriormente é considerado como um *locus* polimórfico multi-alélico (M), e cada haplótipo dentro desse fragmento deve ser representado por um número de 1 a *n*, sendo *n* o número de haplótipos distintos encontrados neste fragmento. É importante enfatizar que um máximo de 49 haplótipos + *missing regions* (-1) pode estar presente em um determinado fragmento. Para atribuir a numeração aos haplótipos de cada fragmento (e eventualmente atribuir uma designação hierárquica definitiva, composta por linhagens e haplótipos derivados), é interessante avaliar a relação molecular entre os haplótipos e suas

frequências por meio de uma *network*. Para isso podem ser usados os *softwares* Network (www.fluxus-engineering.com) ou PopART (<http://popart.otago.ac.nz>), por exemplo.

A inferência dos haplótipos estendidos deve ser feita utilizando os mesmos parâmetros citados anteriormente. Terminada a análise, devem ser aceitos como corretos os haplótipos inferidos com probabilidade superior a 80%. Nesse caso é importante verificar se os haplótipos estendidos inferidos em indivíduos que apresentavam *missing regions* em algum dos fragmentos são compatíveis com os genótipos originalmente determinados em tais regiões: se for inferido um haplótipo que envolve alelos inexistentes na *missing region* desse indivíduo, a inferência deve ser descartada.

3. DESEQUILÍBRIO DE LIGAÇÃO

O arquivo VCF deve ser convertido para o formato ARP através de *script* ferramenta VCFx (versão 0.10.1), disponível em http://www.castelli-lab.net/apps/apps_vcfx.php.

O desequilíbrio de ligação entre os *loci* estudados pode ser avaliado por meio do teste de razão de verossimilhança (Excoffier & Slatkin, 1998) implementado no programa Arlequin 3.5 (Excoffier & Lischer, 2010). Após a reconstrução dos haplótipos, com a fase gamética conhecida, o programa permite estimar os coeficientes de desequilíbrio de ligação entre pares de *loci* (D , D' e r^2).

O coeficiente clássico de desequilíbrio de ligação (D) mede o desvio da associação aleatória entre alelos nos diferentes *loci* (Lewontin & Kojima, 1960), sendo expresso por $D = p_{ij} - p_i p_j$, onde p_{ij} é a frequência do haplótipo formado pelos alelos i no primeiro *locus* e j no segundo e $p_i p_j$ corresponde ao produto das frequências dos alelos i e j .

Já (D') é o coeficiente de desequilíbrio de ligação (D) padronizado pelo máximo valor que esse pode assumir (D_{max}) de acordo com as frequências alélicas (Lewontin, 1964):

$$D'_{ij} = \frac{D_{ij}}{D_{ij\max}}$$

Uma outra medida do desequilíbrio de ligação é representada pelo quadrado do coeficiente de correlação entre as frequências alélicas (r^2), que é expresso em função da medida D :

$$r^2 = \frac{D^2}{p_i(1-p_i) \cdot p_j(1-p_j)}$$

Para visualização gráfica dos blocos haplotípicos presentes ao longo das regiões analisadas e do desequilíbrio de ligação associado pode ser utilizado o programa *Haploview* 4.2 (Barrett, Fry, Maller, & Daly, 2005). Através dessa representação é possível então avaliar o nível de desequilíbrio presente entre cada par de SNP do segmento avaliado.

Os pares de *loci* são considerados em forte DL se o limite superior do IC95% de D' for igual ou superior a 0,98 (o que é consistente com ausência de recombinação histórica) e o limite inferior do intervalo for maior que 0,7. Assim, um bloco haplotípico é definido como uma região onde uma pequena parcela das comparações entre os pares de sítios bialélicos (<5%) apresenta forte evidência de recombinação histórica. Estes 5% são tolerados uma vez que outras forças evolutivas (como mutação recorrente, conversão gênica, erros de montagem do genoma ou de genotipagem), além da recombinação, podem romper os padrões de haplótipo. O método então ordena por tamanho a lista dos possíveis blocos de haplótipos, sendo cada bloco adicionado de forma a não haver sobreposição com o bloco anterior (Gabriel, 2002).

4. FREQUÊNCIAS ALÉLICAS, EQUILÍBRIO DE HARDY-WEINBERG E HETEROZIGOSE ESPERADA

As frequências alélicas para cada sítio de variação (tanto estratificadas por categoria fenotípica quanto globalmente) e a aderência das frequências genotípicas ao modelo de Equilíbrio de Hardy-Weinberg (EHW) podem ser verificadas através do programa GENEPOP v.4.0.10 (Raymond & Rousset, 1995; Rousset, 2008). A heterozigose esperada pode ser calculada a partir do programa Arlequin 3.5 (Excoffier & Lischer, 2010).

5. ASSOCIAÇÕES GENÓTIPO-FENÓTIPO

O teste exato de Fisher bi-caudal, implementado em planilha desenvolvida para esta finalidade (Mendes-Junior C.T.), a qual se utiliza de suplemento disponível em <http://www.obertfamily.com/software/fisherexact.html>, pode ser utilizado para avaliar a existência de associação dos genótipos/alelos/haplótipos com determinado fenótipo.

A mesma planilha realiza ainda o cálculo de *Odds Ratio* (OR) e do intervalo de confiança associado (IC) (Le, 2003; Van Belle, Fisher, Heagerty, & Lumley, 2004). Deve ser adotado o IC de 95% e o limite de significância deve ser estabelecido em $\alpha = 0,05$. O valor de OR permite verificar quantas vezes o caráter em questão (fenótipo avaliado) é mais frequente no grupo de indivíduos apresentando determinado genótipo/alelo/haplótipo, comparativamente ao risco em indivíduos não portador deste fator (Rumel, 1986).

6. REFERÊNCIAS

- Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2), 263–265. <https://doi.org/10.1093/bioinformatics/bth457>
- 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>
- Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3), 564–567. <https://doi.org/10.1111/j.1755-0998.2010.02847.x>
- Excoffier, L., & Slatkin, M. (1998). Incorporating genotypes of relatives into a test of linkage disequilibrium. *American Journal of Human Genetics*, 62(1), 171–180. <https://doi.org/10.1086/301674>

- Gabriel, S. B. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science*, 296(5576), 2225–2229. <https://doi.org/10.1126/science.1069424>
- Le, C. T. (2003). *Introductory Biostatistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/0471308889>
- Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1), 49–67. <https://doi.org/10.1038/nature08185>
- Lewontin, R. C., & Kojima, K. (1960). The Evolutionary Dynamics of Complex Polymorphisms. *Evolution*, 14(4), 458. <https://doi.org/10.2307/2405995>
- 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>
- Raymond, M., & Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *J. Heredity*, 86, 248–249.
- Rousset, F. (2008). GENEPOP'007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, 8(1), 103–106. <https://doi.org/10.1111/j.1471-8286.2007.01931.x>
- Rumel, D. (1986). “Odds Ratio”: Algumas Considerações. *Rev Saúde Publica*, 20(3), 253–258. <https://doi.org/10.1590/S0034-89101986000300011>
- Stephens, M., & Donnelly, P. (2003). Report A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *Am. J. Hum. Genet*, 73(2002), 1162–1169. <https://doi.org/10.1086/379378>
- Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68(4), 978–989. <https://doi.org/10.1086/319501>
- Van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004). *Biostatistics: A Methodology for the Health Sciences*. Hoboken, NJ, USA: John Wiley & Sons, Inc. <https://doi.org/10.1002/0471602396>
- 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>

Anexo E – Protocolo para identificação computacional de interações entre miRNAs e região 3'UTR

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

Protocolo para identificação computacional de interações entre miRNAs e região 3'UTR

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

Fracasso NCA, Silva GV, Castelli EC, 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.

MicroRNAs (miRNAs) atuam na regulação de diversos processos biológicos através da interação entre sua região *seed* e sequências presentes principalmente na região codificadora e 3'UTR dos mRNAs alvo (Mione & Bosserhoff, 2015). Essa interação, dependendo da complementariedade de sequência entre o miRNA e seu alvo, pode resultar na inibição da tradução do mRNA (no caso de complementariedade incompleta) ou facilitar a degradação do mRNA (havendo complementariedade completa) (Huntzinger & Izaurralde, 2011).

Características específicas dos miRNAs (como possibilidade de pareamento incompleto e pequeno tamanho da sequência – 19-25 nucleotídeos) permitem que essas moléculas regulem centenas de mRNAs, tornando bastante complexo o entendimento desta interação (Di Leva & Croce, 2013). Além disso, é necessário considerar que esta regulação pode ainda ser influenciada por variações gênicas presentes nos sítios de mRNA alvo (Palmero et al., 2011). Assim, a compreensão do papel desempenhado por essas moléculas

perpassa a identificação dos seus alvos no genoma. Essa identificação pode ser realizada experimentalmente. Todavia, tendo em vista a complexidade e os custos associados ao método experimental, tem-se buscado alternativas computacionais *in silico* para o reconhecimento preditivo da ligação miRNA-mRNA-alvo (Marín & Vaníek, 2011).

Na predição computacional, utiliza-se as sequências de miRNAs disponíveis em bancos de dados para buscar possíveis alvos, neste caso, sequências 3'UTR do mRNA de determinado gene. A restrição da análise à 3'UTR se deve ao fato de a maioria dos miRNAs se ligarem preferencialmente nessa região (Lewis, Shih, Jones-Rhoades, Bartel, & Burge, 2003).

A fim de identificar miRNAs que se liguem a sequências polimórficas da 3'UTR de um determinado gene, são selecionados os haplótipos encontrados com frequência acima de 1% na amostra populacional.

Tendo em vista o grande número de falso-positivos obtidos através dos métodos *in silico*, optamos por utilizar três algoritmos distintos para predição de miRNAs com afinidade a sequências da 3'UTR: RNAhybrid (Rehmsmeier, Steffen, Höchsmann, Giegerich, & Ho, 2004), miRanda (John et al., 2004) e intaRNA (Busch, Richter, & Backofen, 2008).

Considerando que o miRNA se hibridize ao alvo de maneira energeticamente ótima, o RNAhybrid se baseia no cálculo da energia mínima livre (MFE) de hibridização entre as posições iniciais (*seed*) no miRNA e no alvo. O miRanda também se baseia no cálculo da MFE de hibridização para predição de miRNAs afins, mas utiliza ainda informações relacionadas à conservação evolutiva da sequência alvo e à posição do alvo em 3'UTRs alinhadas de genes homólogos. Por sua vez, o intaRNA se baseia em um *score* combinado da energia de interação, que é calculado como a soma da energia livre de hibridização e da energia livre necessária para tornar os sítios de interação acessíveis.

Para realizar uma avaliação conjunta utilizando esses três algoritmos é empregado um *software*, mirhunt (Porto et al., 2015) (http://www.castelli-lab.net/apps/apps_mirhunt.php), que automatiza a análise como brevemente descrito: a) as sequências de mRNA da 3'UTR são fragmentadas em uma série de subsequências de 500 nucleotídeos com uma sobreposição de 50 nucleotídeos iniciais e finais; b) os três algoritmos são utilizados para prever interações miRNA/mRNA considerando cada uma das subsequências; c) os resultados gerados apresentam todas as interações miRNA/mRNA encontradas por cada algoritmo; d) mantem-se

nos resultados somente a ligação mais estável nos casos em que são identificadas mais de uma interação para um mesmo miRNA com afinidade para uma mesma região-alvo (para reconhecimento de uma mesma região alvo, os últimos nucleotídeos pareados da região *seed* são utilizados como referência, com uma tolerância máxima de 3 nucleotídeos); e) os resultados encontrados pelos três algoritmos são comparados; f) um sistema de pontuação é aplicado para cada interação miRNA/alvo. Os parâmetros utilizados para o miRanda são: -4 para extensão e -9 para abertura do *gap*, limiar de pontuação estabelecido como 50 e mínima energia livre configurada para -1 kcal/mol. Para o RNAhybrid: mínima energia livre de ligação configurada para -1 kcal/mol e número de *hits* ajustado para 10. O intaRNA é um método livre de parâmetro, então a configuração padrão é utilizada (Porto et al., 2015).

O banco de dados contendo os miRNAs humanos até então identificados (2.588) pode ser obtido a partir do miRBase, release 21 de junho de 2014 (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006). Versão mais recente, caso disponível, deve ser utilizada.

Os algoritmos são utilizados para inferir a melhor interação miRNA/mRNA (menor valor de MFE) para cada miRNA e cada sequência da 3'UTR disponível para análise. Finalmente, um sistema de pontuação é aplicado para cada conjunto de dados, como descrito detalhadamente em Porto et al. (2015). Essas pontuações têm como objetivo identificar as interações mais estáveis e/ou específicas entre miRNAs e seus alvos, permitindo a seleção de miRNAs mais influentes, evitando o uso de limites arbitrários. Ao compor os resultados dos três métodos, interações falso-positivas eventualmente identificadas por apenas um dos métodos são descartadas por não resultarem em uma alta pontuação.

A primeira pontuação (denominada valor ε) reflete a estabilidade da interação miRNA/mRNA. O menor valor ε encontrado por cada algoritmo é utilizado como referência do algoritmo para propósito de normalização. Este procedimento tem como objetivo tornar os MFEs dos diferentes algoritmos comparáveis. A pontuação de ε para cada interação miRNA/mRNA é calculada a partir da razão entre o MFE daquela interação e aquele da referência. Após a normalização, a pontuação de energia composta para cada interação miRNA/mRNA é obtida pela soma das pontuações de cada algoritmo: $\varepsilon = \varepsilon_{\text{RNAhybrid}} + \varepsilon_{\text{miRanda}} + \varepsilon_{\text{intaRNA}}$. Assim, quanto maior for a pontuação ε maior a probabilidade de ocorrer uma forte interação entre o miRNA e o alvo.

A segunda pontuação (denominada valor s) reflete a especificidade da interação miRNA/mRNA, comparando as interações observadas com uma distribuição nula que leva em consideração a interação destes 2588 miRNAs em 23.464 sequências 3'UTR de genes humanos. Quanto maior for a pontuação s , maior a probabilidade do miRNA em questão interagir somente com o gene estudado.

A terceira pontuação (denominada valor ϵs) é calculada como um valor combinado que reflete ambas as pontuações, ϵ e s . Quanto maior a pontuação ϵs , maior a probabilidade de que o miRNA em questão interaja de forma forte e específica com o gene estudado.

REFERÊNCIAS

- Busch, A., Richter, A. S., & Backofen, R. (2008). IntaRNA: Efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24), 2849–2856. <https://doi.org/10.1093/bioinformatics/btn544>
- Di Leva, G., & Croce, C. M. (2013). MiRNA profiling of cancer. *Current Opinion in Genetics and Development*, 23(1), 3–11. <https://doi.org/10.1016/j.gde.2013.01.004>
- Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue), D140–4. <https://doi.org/10.1093/nar/gkj112>
- Huntzinger, E., & Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews. Genetics*, 12(2), 99–110. <https://doi.org/10.1038/nrg2936>
- John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2004). Human MicroRNA targets. *PLoS Biology*, 2(11), e363. <https://doi.org/10.1371/journal.pbio.0020363>
- Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). 33- Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7), 787–798. [https://doi.org/10.1016/S0092-8674\(03\)01018-3](https://doi.org/10.1016/S0092-8674(03)01018-3)
- Marín, R. M., & Vaníek, J. (2011). Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Research*, 39(1), 19–29. <https://doi.org/10.1093/nar/gkq768>

- Mione, M., & Bosserhoff, A. (2015). MicroRNAs in melanocyte and melanoma biology. *Pigment Cell and Melanoma Research*, 28(3), 340–354. <https://doi.org/10.1111/pcmr.12346>
- Palmero, E. I., de Campos, S. G. P., Campos, M., de Souza, N. C. N., Guerreiro, I. D. C., Carvalho, A. L., & Marques, M. M. C. (2011). Mechanisms and role of microRNA deregulation in cancer onset and progression. *Genetics and Molecular Biology*, 34(3), 363–370. <https://doi.org/10.1590/S1415-47572011000300001>
- Porto, I. O. P., Mendes-Junior, C. T., Felício, L. P., Georg, R. C., Moreau, P., Donadi, E. A., ... Castelli, E. C. (2015). MicroRNAs targeting the immunomodulatory HLA-G gene: A new survey searching for microRNAs with potential to regulate HLA-G. *Molecular Immunology*, 65(2), 230–241. <https://doi.org/10.1016/j.molimm.2015.01.030>
- Rehmsmeier, M., Steffen, P., Höchsmann, M., Giegerich, R., & Ho, M. (2004). Fast and effective prediction of microRNA / target duplexes. *Spring*, (2003), 1507–1517. <https://doi.org/10.1261/rna.5248604.and>