UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
DEPARTAMENTO DE GENÉTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

MARIA LUIZA GUIMARÃES DE OLIVEIRA

**Sequenciamento de nova geração do gene *IRF4*: identificação de variações
associadas a fenótipos de pigmentação na população brasileira**

Ribeirão Preto - SP

2016

MARIA LUIZA GUIMARÃES DE OLIVEIRA

Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira

Dissertação apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências, área de concentração Genética

Orientador: Prof. Dr. Celso Teixeira Mendes Junior

Ribeirão Preto – SP

2016

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.

Oliveira, Maria Luiza Guimarães de

Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira / Maria Luiza Guimarães de Oliveira; orientador: Prof. Dr. Celso Teixeira Mendes Junior. - Ribeirão Preto, 2016.

151p; 30cm

Dissertação de Mestrado apresentada à Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo para obtenção do título de Mestre em Ciências, área de concentração Genética – USP/ FMRP/ Departamento de Genética.

1- diversidade genética; 2- fator regulador de interferon-4; 3- sequenciamento de nova geração; 4 – pigmentação.

Nome: Oliveira, Maria Luiza Guimarães de

Título: Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira

Dissertação apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências, área de concentração Genética

Aprovado em:

Banca Examinadora

Prof. Dr. _____ Instituição _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição _____

Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição _____

Julgamento: _____ Assinatura: _____

A minha família!

AGRADECIMENTOS

A minha família, que mesmo distante nunca deixou de se fazer presente. Aos meus pais, por acreditarem em mim e pelo esforço que fazem ao apoiarem minhas decisões. Aos meus amados irmãos, Edgar e Osmar, por sempre me incentivarem a seguir em frente. A todos da minha família, aos Guimarães e aos Oliveira, por não me deixarem nunca sentir desamparada. Amo muito todos vocês!

Ao meu orientador, Prof. Dr. Celso Teixeira Mendes Junior, por me aceitar no laboratório, pelo exemplo, pela imensa compreensão, paciência, incentivo e pelos inúmeros conhecimentos compartilhados.

Ao Prof. Dr. Erick da Cruz Castelli, por todo o conhecimento compartilhado, pelo imprescindível auxílio prestado nas análises de bioinformática e pela disposição em ajudar na resolução dos problemas que surgiram pelo caminho.

A todos os colegas do Laboratório de Pesquisas Forenses e Genômicas: Alison, Nadia, Letícia, Guilherme, Edilene e ao técnico do laboratório, André. Por toda ajuda prestada e pelos momentos compartilhados, seja fazendo bibliotecas ou nas reuniões no Marcão. Em especial ao querido Alison, que incansavelmente me ajudou com as análises, auxiliando quase até o Natal na resolução de problemas com os programas ou formato dos arquivos. Muito obrigada por tudo, mesmo!

Às queridas Pati e Aline pela amizade e por tudo que me ensinaram. Exemplos de superação e inspiração a quem sou muito grata!

Aos amigos: aos de Santa Catarina, às amigas da Bio, ao pessoal parceiro do Rio. Obrigada pela amizade e por todos os ótimos momentos. Em especial, à Mari, por todo o carinho que tem por mim e que é recíproco, por estar sempre ali, seja por mensagem, áudio ou ao vivo. É uma benção te ter como amiga! À Amandita, pela amizade ao longo desses anos de Bio e que segue adiante. Muitas saudades da sua cia, sempre presente!

Aos amigos que fiz em Ribeirão, não muitos, mas imprescindíveis! À Denyse, pela amizade de sempre e por ser alguém com quem sempre posso contar! Por ter me acolhido no grupo da Biocel, por me incentivar da forma mais sincera possível, pelo carinho e apoio nos momentos em que se tornou mais difícil estar aqui. À Nadia, que me acolheu em um momento muito difícil. Pela amizade, pelos conselhos, pela paciência incomensurável em me ouvir. Não tenho palavras para agradecer! São nos momentos mais difíceis que aparecem anjos na nossa vida, que nos dão força para seguir em frente. Obrigada por ser um deles!

À Universidade de São Paulo, pela infraestrutura cedida.

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

Aos professores e à Coordenação do Programa de Pós-Graduação em Genética, pelos ensinamentos compartilhados e apoio prestado. Em especial, à Susie, pela dedicação e disposição em ajudar sempre que necessário.

Ao Prof. Dr. Aguinaldo Luiz Simões, por gentilmente ceder à utilização dos equipamentos de seu laboratório.

À Cláudia, Elisabete e Rubens pelo auxílio prestado na realização de técnicas imprescindíveis a este trabalho.

Ao Laboratório de Imunopatologia Molecular, em especial ao Prof. Dr. Eduardo Antônio Donadi, pelo suporte e estrutura para realização deste trabalho e à Juliana Doblas Massaro, Sandra e Flávia, que auxiliaram na realização dos ensaios de sequenciamento.

Aos professores membros da banca examinadora, que gentilmente aceitaram participar e contribuir com este trabalho.

Aos voluntários que doaram uma parte de seu tempo e história. Agradeço imensamente a cada um.

A Deus!

Muito obrigada a todos!!

RESUMO

OLIVEIRA, M. L. G. **Sequenciamento de nova geração do gene *IRF4*: identificação de variações associadas a fenótipos de pigmentação na população brasileira**. 2016. 151 páginas. Dissertação de Mestrado – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, 2016.

O gene fator regulador de interferon 4 (*IRF4*), localizado na região cromossômica 6p25-p23, é um membro da família de fatores reguladores de interferon (IRF), um grupo de fatores de transcrição de ligação ao DNA, sendo *IRF4* primariamente associado ao desenvolvimento e resposta imune e expresso exclusivamente em células do sistema imunológico e em linhagens melanocíticas. Embora muitos estudos tenham associado *IRF4* a diversas condições, como melanoma e leucemia linfocítica crônica, um recente *Genome-Wide Association Study* (GWAS) identificou que alelos do SNP rs12203592 (intron 4) estão associados com variação fenotípica em relação à presença de sardas, pigmentação da pele, cabelos e olhos. Estudos funcionais realizados em células melanocíticas humanas e de camundongos revelaram que este SNP está diretamente envolvido na regulação da expressão de *IRF4*, sugerindo uma clara função na pigmentação do melanócito. Apesar destes achados, a diversidade das regiões regulatórias e codificadora de *IRF4* não foi até o momento analisada em populações miscigenadas. A fim de avaliar se outros sítios de variação ao longo do gene *IRF4* podem estar associados à pigmentação humana, as regiões regulatórias (promotora e 3'UTR) e codificadora (9 exons e regiões intrônicas flanqueadoras, incluindo o SNP rs12203592) foram analisadas por sequenciamento de nova geração em uma amostra miscigenada da população brasileira. A amostra populacional foi composta por 228 indivíduos não aparentados de Ribeirão Preto, estado de São Paulo, Brasil, os quais foram estratificados de acordo com a pigmentação da pele (clara, média e escura), olhos (azul, verde, castanho-claros e castanho-escuros), cabelo (ruivo, loiro-claro, loiro-escuro, castanho-claro, castanho-escuro e preto) bem como em relação à presença de sardas e intensidade de cabelos grisalhos. Bibliotecas de DNA foram preparadas utilizando o Sistema de Enriquecimento de Alvo *Haloplex* (*Agilent Technologies*) e sequenciadas na plataforma *MiSeq* (*Illumina*). Os pacotes de *software* CutAdapt, BWA and GATK foram utilizados, respectivamente, para trimagem das sequências dos adaptadores, alinhamento e identificação de variantes. Haplótipos e alelos não identificados foram inferidos pelo método PHASE, embora a fase conhecida entre os

sítios de variação (obtida pelo GATK) tenha sido levada em consideração. Um total de 105 sítios de variação foram identificados. Apenas dois deles apresentaram frequências genotípicas que não atendem ao esperado pelo equilíbrio de Hardy-Weinberg (EHW). Dezoito destes SNPs apresentaram forte associação a pelo menos uma característica de pigmentação. Entretanto, se a conservadora correção de Bonferroni para múltiplos testes for levada em consideração, apenas duas associações, ambas envolvendo o SNP rs12203592, permanecem significativas: a associação do alelo T com pele clara e olhos azuis. Este resultado está de acordo com estudos prévios, que reportam que o alelo rs12203592*T leva a uma menor ativação de IRF4 e a uma expressão reduzida da tirosinase, resultando em sensibilidade ao sol e olhos azuis. Foi inferido um total de 101 haplótipos, estando a distribuição destes de acordo com o esperado pelo EHW. Quando os haplótipos foram divididos em haplótipos da promotora, codificadora e 3'UTR foram observadas, respectivamente, 17, 29 e 37 diferentes combinações haplotípicas. Várias associações foram identificadas, particularmente envolvendo o haplótipo mais frequente da promotora, os dois haplótipos mais frequentes da codificadora e o haplótipo mais frequente da 3'UTR, todos associados com pele clara, olhos azuis, cabelos castanhos e cabelos grisalhos. Estes resultados sugerem que outras variantes além de rs12203592, quando consideradas em um contexto haplotípico, são associadas com a pigmentação humana.

Palavras-chave: Fator Regulador de Interferon 4, rs12203592, diversidade genética, Sequenciamento de Nova Geração, Brasil.

ABSTRACT

OLIVEIRA, M. L. G. **Next Generation Sequencing of gene *IRF4*: identification of variations associated with pigmentation traits in the Brazilian population.** 2016. 151 pages. Dissertation (Master's Degree) – Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, São Paulo, Brazil.

The Interferon Regulatory Factor 4 (*IRF4*) gene, located at chromosomal region 6p25-p23, is a member of the interferon regulatory factor (IRF) family, a group of DNA-binding transcription factors, with the IRF4 primarily associated with immune system development and response and expressed exclusively in immune system cells and melanocytic lineages. Although many studies have shown that IRF4 is associated with many human conditions, such as melanoma and chronic lymphocytic leukemia, a recent Genome-Wide Association Study (GWAS) identified that alleles from the SNP rs12203592 (intron 4) is also associated with phenotypic variation regarding presence of freckles, hair, eye and skin pigmentation. Functional studies in human and mice melanin-containing cells revealed that such SNP is directly involved in the regulation of IRF4 expression, suggesting a clear role in melanocyte pigmentation. In spite of these findings, the regulatory and coding IRF4 diversities in admixed populations have not been evaluated so far. In order to verify if other variation sites spread across the IRF4 gene may be associated with human pigmentation, the regulatory (promoter and 3'UTR regions) and coding (9 exons and flanking intronic regions, including the SNP rs12203592) regions were analyzed by next-generation sequencing procedures in a Brazilian admixed population sample. The population sample was composed of 228 unrelated individuals from the Ribeirão Preto area, São Paulo State, Brazil, which were stratified according to eye (blue, green, hazel, light-brown, and dark-brown), hair (red, blond, dark-blond, light-brown, dark-brown and black) and skin (light, intermediate and dark) pigmentation, as well as regarding the presence of freckles and intensity of hair greying. DNA libraries were prepared using the Haloplex Target Enrichment System (Agilent Technologies) and sequenced at the MiSeq platform (Illumina). CutAdapt, BWA and GATK software packages were used for trimming adaptor sequences, alignment and genotype calling, respectively. Missing alleles and haplotypes were inferred by using the PHASE method, although the known phase between variable sites (obtained by GATK) was taken into account. A total of 105 variation sites were identified. Only two of them presented genotype frequencies that did not fit Hardy-

Weinberg equilibrium (HWE) expectations. Eighteen of these SNPs presented strong association with at least one pigmentation feature. However, if the conservative Bonferroni correction for multiple tests is taken into account, only two associations, both of them involving the rs12203592 SNP, remain significant: allele T associated with light skin and blue eyes. This result is in agreement with previous reports that the rs12203592*T allele leads to reduced IRF4 activation and reduced tyrosinase expression, leading to sun sensitivity and blue eyes. A total of 101 different haplotypes were inferred, and haplotype distribution was in agreement to HWE expectations. When haplotypes were subdivided in promoter, coding and 3'UTR haplotypes, 17, 29 and 37 different haplotypes were observed, respectively. Various associations were identified, particularly involving the most frequent promoter haplotype, the two most frequent coding (only one of them with allele rs12203592*T), and the most frequent 3'UTR, all of them with light skin, blue eyes, brown hair and hair greying. These results suggest that other variation sites besides rs12203592, when considered in a haplotypic background, are associated with human pigmentation.

Keywords: Interferon Regulatory Factor 4, rs12203592, genetic diversity, Next Generation Sequencing, Brazil.

LISTA DE TABELAS (151 páginas)

Tabela 1: Descrição dos componentes da linha de comando executada no CutAdapt...	52
Tabela 2: Descrição dos componentes da linha de comando executada no BWA.....	53
Tabela 3: Descrição dos componentes de um arquivo SAM.....	54
Tabela 4: Campos obrigatórios de uma linha da seção de alinhamento de um arquivo SAM.	55
Tabela 5: Descrição dos componentes da linha de comando executada no GATK	59
Tabela 6: Descrição dos componentes da linha de comando executada no ReadBackedPhasing.....	61
Tabela 7: Descrição dos componentes da seção de cabeçalho de um arquivo VCF.....	62
Tabela 8: Descrição dos componentes do corpo de um arquivo VCF	63
Tabela 9: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de pele e presença de sardas.....	75
Tabela 10: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de olhos	78
Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo	81
Tabela 12: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de presença de cabelos brancos	87
Tabela 13: Alelos e genótipos de SNPs/indels do gene <i>IRF4</i> associados a características de pigmentação	91
Tabela 14: Haplótipos formados por SNPs/indels do gene <i>IRF4</i> associados a características de pigmentação	92
Tabela 15: Haplótipos formados por SNPs de <i>IRF4</i> associados a características de pigmentação	93
Tabela 16: Pontos de variação presentes ao longo dos haplótipos da região promotora que apresentaram frequência $\geq 1\%$ na amostra populacional estudada	96
Tabela 17: Pontos de variação presentes ao longo dos haplótipos da região codificadora que apresentaram frequência $\geq 1\%$ na amostra populacional estudada	97
Tabela 18: Pontos de variação presentes ao longo dos haplótipos da CDS, que apresentaram frequência $\geq 1\%$ na amostra populacional estudada	98

Tabela 19: microRNAs com alvo para o gene <i>IRF4</i> apresentando os 20 maiores valores de ΔE^* , considerando as UTRs com frequência de pelo menos 1% na amostra estudada.....	99
Tabela 20: microRNAs com alvo para o gene <i>IRF4</i> apresentando os 20 maiores valores de ΔS^* , considerando as UTRs com frequência de pelo menos 1% na amostra estudada.....	100
Tabela 21: microRNAs com alvo para o gene <i>IRF4</i> apresentando os 20 maiores valores de ΔES^* , considerando as UTRs com frequência de pelo menos 1% na amostra estudada	101
Tabela 22: microRNAs selecionados por apresentaram os 20 maiores valores de $\Delta E, S$ e ES , distribuídos conforme sua presença em cada um dos índices considerados	102
Tabela 23: Pontos de variação presentes ao longo dos haplótipos da 3'UTR que apresentaram frequência $\geq 1\%$ na amostra populacional estudada	103
Tabela 24: Comparação das médias do índice de melanina em cada grupo de indivíduos, classificados de acordo com a escala de Fitzpatrick.....	104
Tabela 25: Comparação das médias do índice de melanina em cada grupo de indivíduos, classificados de acordo com o IBGE	105
Tabela 26: Análise da associação entre os índices de melanina em cada grupo de genótipos, considerando os 18 SNPs associados a alguma das características de pigmentação avaliadas	106
Tabela 27: Frequências alélicas dos SNPs localizados na região promotora, associados a algum dos fenótipos de pigmentação avaliados nas populações que compõem o Projeto 1000 <i>Genomes</i> e na população do presente estudo.....	116
Tabela 28: Frequências alélicas dos SNPs localizados na região codificadora, associados a algum dos fenótipos de pigmentação avaliados nas populações que compõem o Projeto 1000 <i>Genomes</i> e na população do presente estudo.....	122
Tabela 29: Frequências alélicas dos SNPs localizados na região 3'UTR do gene <i>IRF4</i> , associados a algum dos fenótipos de pigmentação avaliados nas populações que compõem o Projeto 1000 <i>Genomes</i> e na população do presente estudo.....	127

LISTA DE FIGURAS (151 páginas)

Figura 1: Fases do desenvolvimento dos melanossomos.	22
Figura 2: Corte histológico representando a estrutura da pele humana.	23
Figura 3: Representação esquemática da rota de biossíntese da melanina.....	24
Figura 4: Esquema da estrutura da pele humana: pigmentação clara e escura.	25
Figura 5: Diferenças observadas de acordo com a coloração dos olhos.....	26
Figura 6: Modelo da relação entre MITF, TFAP2 α e IRF4 na expressão de TYR	31
Figura 7: Análise dos potenciais sítios de ligação para fatores de transcrição no gene <i>IRF4</i>	32
Figura 8: Perfil de picos esperado no eletroferograma para as bibliotecas de fragmento deste ensaio	47
Figura 9: Gráfico de distribuição da qualidade por base no programa FastQC.	50
Figura 10: Representação dos alinhamentos das <i>reads</i> na região do gene <i>IRF4</i> no programa IGV_2.3.67.	57
Figura 11: Porcentagem de reads com qualidade igual ou superior a 30	72
Figura 12: Distribuição dos indivíduos quanto às características fenotípicas observadas	74
Figura 13: D' e frequências dos haplótipos inferidos em cada bloco pelo Haploview.....	94
Figura 14: Representação do Desequilíbrio de ligação entre os SNPs presentes ao longo da região gênica avaliada através dos valores de D' e LOD <i>score</i>	95
Figura 15: Representação gráfica das informações disponibilizadas no ENCODE para a região do gene <i>IRF4</i>	115

SUMÁRIO

1. INTRODUÇÃO	19
1.1 IDENTIFICAÇÃO HUMANA COM FINS FORENSES	20
1.2 BIOLOGIA DA PIGMENTAÇÃO HUMANA.....	22
1.3 DIVERSIDADE DA POPULAÇÃO BRASILEIRA	27
1.4 MARCADORES GENÉTICOS NA PREDIÇÃO DO FENÓTIPO DE PIGMENTAÇÃO.....	28
1.5 O GENE <i>IRF4</i>	29
2. JUSTIFICATIVA	34
3. HIPÓTESE.....	36
4. OBJETIVOS	38
5. MATERIAL E MÉTODOS	40
5.1 COLETA DAS AMOSTRAS	41
5.2 ANÁLISE LABORATORIAL.....	42
5.2.1 Extração do DNA genômico	42
5.2.2 Preparo dos reagentes para enriquecimento de alvo	42
5.2.3 Enriquecimento do alvo e preparo da biblioteca.....	43
5.2.3.1 Digestão e hibridização do DNA	44
5.2.3.2 Captura, ligação, PCR e eluição.....	45
5.2.3.3 Purificação da biblioteca	45
5.2.4 Controle de qualidade e quantificação da biblioteca.....	46
5.2.5 Sequenciamento	47
5.3 ANÁLISE DE DADOS	48
5.3.1 Análise primária	48
5.3.2 Análise secundária	49
5.3.2.1 Avaliação da qualidade.....	50

5.3.2.2 Trimagem.....	51
5.3.2.3 Alinhamento ao genoma.....	52
5.3.2.3.1 SAM.....	54
5.3.2.3.2 BAM	55
5.3.3 Análise terciária	57
5.3.3.1 Determinação de variantes.....	57
5.3.3.2 Remoção de genótipos de baixa qualidade.....	60
5.3.3.3 Faseamento experimental	61
5.3.3.4 VCF	62
5.4 ANÁLISES ESTATÍSTICAS	64
5.4.1 Faseamento computacional	64
5.4.2 Desequilíbrio de ligação.....	66
5.4.3 Frequências alélicas, equilíbrio de Hardy-Weinberg e heterozigose esperada	67
5.4.4 Associações Genótipo-Fenótipo.....	67
5.5 ANÁLISES DE MICRORNAS	68
<u>6. RESULTADOS</u>	<u>71</u>
6.1 SEQUENCIAMENTO DE NOVA GERAÇÃO.....	72
6.2 AMOSTRA POPULACIONAL.....	72
6.3 FREQUÊNCIAS ALÉLICAS, HAPLOTÍPICAS E HETEROZIGOSE	74
6.4 ASSOCIAÇÕES ENCONTRADAS.....	91
6.5 DESEQUILÍBRIO DE LIGAÇÃO	94
6.6 ANÁLISE PROMOTORA.....	96
6.7 ANÁLISE DA REGIÃO CODIFICADORA.....	97
6.7.1 ANÁLISE DA REGIÃO <i>CODING DNA SEQUENCE</i> (CDS)	98
6.8 ANÁLISE DE MICRORNAS	99
6.9 ANÁLISES DOS DADOS DE ÍNDICE DE MELANINA	104

7. DISCUSSÃO	107
7.1 AMOSTRA POPULACIONAL.....	108
7.2 DESVIOS EM RELAÇÃO AO EHW.....	109
7.3 ASSOCIAÇÕES ENCONTRADAS	110
7.3.1 Associações de SNPs localizados na região promotora	111
7.1.3 Associação de SNPs encontrados em introns da região codificadora	118
7.1.4 Associação de SNPs encontrados na CDS	124
7.1.5 Associação de SNPs encontrados na região 3'UTR.....	125
7.1.5.1 MicroRNAs preditos considerando a região 3'UTR de <i>IRF4</i>	130
7.1.5 Associação de SNPs encontrados na região intergênica	132
8. CONCLUSÕES.....	134
9.REFERÊNCIAS BIBLIOGRÁFICAS	136
10.ANEXOS	147

1. INTRODUÇÃO

1.1 IDENTIFICAÇÃO HUMANA COM FINS FORENSES

A descoberta por Allec Jeffreys de um padrão extremamente variável e herdável de sequências repetidas no DNA tornou possível traçar o perfil genético de um indivíduo, o que ficou conhecido como “DNA *fingerprinting*” ou impressão digital do DNA (JEFFREYS; WILSON, 1985). Utilizando a análise de VNTRs (*Variable Number of Tandem Repeats*) por RFLP (*Restriction Fragment Length Polymorphism*) foi obtido um alto poder discriminatório, porém, dependente de grande quantidade de DNA disponível de boa qualidade (GOODWIN, 2007).

O advento da PCR (*Polymerase Chain Reaction*) permitiu que quantidades mínimas de DNA, extraídas a partir dos mais diversos tipos de amostras como sangue, saliva, sêmen, pudessem ser genotipadas com sucesso (MULLIS, 1990). Polimorfismos do *locus* HLA-DQA1 foram os primeiros marcadores analisados via PCR (SAIKI et al., 1985). No entanto, este sistema não suplantou o poder discriminatório obtido com o uso dos VNTRs, sendo então substituído pela utilização de *loci* microsatélites ou STRs (*Short Tandem Repeats*) (BUDOWLE; VAN DAAL, 2008).

Os STRs, compostos por curtas sequências repetidas *in tandem*, em geral com unidade de repetição de 4 pb, são extremamente polimórficos e bastante abundantes no genoma humano, o que os torna adequados à identificação forense (EDWARDS et al., 1991). Estes marcadores são mundialmente utilizados e, atualmente, ocupam lugar de destaque no sistema CODIS (*Combined DNA Index System*) do FBI (*Federal Bureau of Investigation*), que estabeleceu um painel inicial de 13 *loci* STRs a serem genotipados para estabelecer um perfil de DNA. Os perfis de DNA obtidos de uma cena de crime podem ser comparados aos perfis de suspeitos ou aos perfis depositados no CODIS por outros laboratórios criminais, permitindo a identificação ou exclusão de suspeitos (FBI, 2016). Contudo, em situações onde são obtidas amostras extremamente degradadas, a utilização de SNPs (*Single Nucleotide Polymorphisms*) torna-se mais adequada.

SNPs representam o tipo mais comum de variação no genoma humano e são na sua maioria bialélicos, o que os torna individualmente menos informativos que os *loci* STRs. Entretanto, oferecem: a possibilidade de utilizar pequenos fragmentos amplificados por PCR (com cerca de 40-80 pb), a facilidade de análise em larga escala e a característica de estabilidade evolutiva, com taxas de mutação bastante reduzidas (BUTLER; SHE; MCCORD, 2003). Essas características tornam estes marcadores

adequados em quaisquer situações onde um perfil de STRs não é informativo ou não pode ser obtido, o que ocorre principalmente em casos onde o suspeito ainda não está identificado, em uma situação de desastre em massa ou no caso de pessoas desaparecidas (BUDOWLE; VAN DAAL, 2008).

Nos casos identificados acima, pode ser utilizada a fenotipagem do DNA (*DNA Phenotyping*), que inclui como abordagens a inferência sobre a ancestralidade biogeográfica e a predição de características físicas visíveis (KAYSER; DE KNIJFF, 2011).

A utilização de AIMs (*Ancestry-Informative Markers*) pode revelar a ancestralidade biogeográfica de um suspeito, direcionando a investigação para indivíduos com determinada aparência. AIMs são polimorfismos que apresentam alelos com frequências bastante distintas em comparações envolvendo diferentes populações mundiais (ROSENBERG et al., 2002; SHRIVER et al., 2003). Diferentes populações são geralmente heterogêneas em termos de ancestralidade genética (CARDON; PALMER, 2003). Porém, a população estudada pode estar subestruturada, ou, dividida em subgrupos geneticamente e fenotipicamente diferenciados. Essa estratificação ocorre devido a padrões preferenciais de reprodução entre determinados grupos da população (seja por causa da presença de barreiras geográficas ou mesmo em função de questões culturais), o que acaba levando a diferenças nas frequências alélicas nessas subpopulações (KAYSER; KNIJFF, 2011). Deste modo, o emprego de AIMs se mostra adequado para um direcionamento inicial das investigações.

A outra abordagem consiste na utilização de SNPs encontrados em genes associados a características físicas visíveis como cor dos olhos, da pele, do cabelo (LIU et al., 2010; SULEM et al., 2007), altura (LANGO ALLEN et al., 2010) e espessura capilar (MEDLAND et al., 2009). Estes marcadores podem guiar de maneira mais acurada a identificação do suspeito de um crime em função das suas características fenotípicas, permitindo uma predição mais precisa do que a proporcionada pelos AIMs, uma vez que a ancestralidade genômica nem sempre está correlacionada com a aparência física (PARRA; KITTLES; SHRIVER, 2004; DURSO et al., 2014). Entretanto, pouco ainda é conhecido a respeito dos fatores genéticos que determinam a aparência humana (KAYSER; DE KNIJFF, 2011).

1.2 BIOLOGIA DA PIGMENTAÇÃO HUMANA

Populações humanas apresentam grande variabilidade fenotípica, estando a pigmentação da pele, olhos e cabelos entre as características mais visíveis dessa variação (SULEM et al., 2007).

A pigmentação da pele é determinada pela deposição de três pigmentos: melanina, hemoglobina e carotenoides. Destes, a melanina é o principal (LIN; FISHER, 2007). A melanina é produzida nos melanossomos, compartimentos especializados do Complexo de Golgi, presentes nos melanócitos. Os melanossomos passam por diferentes estágios de maturação (Figura 1).

Figura 1: Fases do desenvolvimento dos melanossomos. No estágio I os premelanossomos se constituem em vacúolos pequenos e vazios. No estágio II, os melanossomos passam a apresentar uma matriz fibrilar. O início da produção de melanina ocorre no estágio III, quando o pigmento é depositado nas fibras proteicas. No estágio IV, a melanina preenche todo o melanossomo. Adaptado de: Cichorek, et al. 2013.

Quando completamente preenchidos por melanina, os melanossomos são exportados dos melanócitos para os queratinócitos adjacentes (Figura 2). Assim, as diferenças na pigmentação resultam da quantidade, tamanho, composição e distribuição dos melanossomos e da razão entre os tipos de melanina (eumelanina/feomelanina) produzida por estas organelas (LIN; FISHER, 2007).

Figura 2: Corte histológico representando a estrutura da pele humana. A camada superior da pele (epiderme) contém os queratinócitos, caracterizados por apresentar um núcleo grande (corados em azul). A camada logo abaixo, corada em rosa, é a derme. Os melanócitos (indicados pelas setas) estão localizados na camada basal da epiderme, na junção com a derme. Adaptado de: Lin; Fisher, 2007.

A biossíntese da melanina envolve uma série de reações químicas e enzimáticas, sendo essa produção ativada pela ligação do hormônio estimulante de α -melanócitos (α -MSH) ao receptor de melanocortina-1 (MC1R), aumentando o nível de AMP-cíclico (cAMP), o que leva à ativação do fator de transcrição associado a microftalmia (MITF). MITF ativa a transcrição do gene fator regulatório de interferon 4 (*IRF4*). Em conjunto com *IRF4*, MITF estimula a produção da enzima tirosinase (TYR). Ainda que MITF independa de *IRF4* para ativar essa produção, quando ocorre a ativação conjunta *IRF4*-MITF o estímulo é muito mais forte. Além disso, MITF atua também sob a proteína 1 relacionada à tirosinase (TYRP1) e a proteína 2 relacionada à tirosinase (TYRP2) - também conhecida como dopacroma tautomerase (DCT), as quais por sua vez não parecem ser afetadas por *IRF4* (PRAETORIUS et al., 2013b).

TYR promove a conversão do aminoácido tirosina a 3,4-*dihydroxyphenylalanine* (DOPA) e sua subsequente oxidação a DOPAquinona, sendo esse passo limitante ao processo, uma vez que a síntese de ambos os tipos de melanina utiliza DOPAquinona como substrato (COSTIN; HEARING, 2007). Essa produção de melanina pode ocorrer através de duas rotas alternativas (Figura 3):

- na ausência do aminoácido Cisteína, DOPAquinona produz eumelanina (pigmento preto/castanho). Através da atividade catalítica das proteínas TYRP1 e DCT, são sintetizados *dihydroxyindole* (DHI) e *dihydroxyindole-2-carboxylic acid* (DHICA), os quais são convertidos à

indole-5,6-quinone e *indole-5,6-quinone carboxylic acid*, respectivamente, resultando na produção de eumelanina;

- na presença do aminoácido Cisteína, DOPAquinona produz feomelanina (pigmento vermelho/amarelo).

Figura 3: Representação esquemática da rota de biossíntese da melanina. α -MSH estimula MC1R, liberando cAMP, o que ativa MITF, que por sua vez, aciona TYR, TYRP1, TYRP2 e ativa a transcrição de *IRF4*. *IRF4*-MITF promovem então uma ativação conjunta de TYR, mais intensa que aquela promovida por MITF, isoladamente. TYR catalisa a hidroxilação da tirosina a DOPA; a oxidação de DOPA a DOPAquinona. Na ausência de cisteína, a reação adicional culmina na produção de eumelanina. Na presença de cisteína, DOPAquinona é convertida a 5-S-cysteiny DOPA, o que resulta na produção de feomelanina Adaptado de: Marcorin, 2016.

A pele humana apresenta uma mistura dos dois tipos de melanina, sendo proporção de eumelanina presente o que define a pigmentação da pele. Uma vez que a cisteína é um fator limitante para a biossíntese de feomelanina, sua produção é normalmente baixa, não se correlacionando bem com a intensidade de pigmentação da pele (SLOMINSKI et al., 2004; CICHOREK et al., 2013). A distribuição dos melanossomos também varia de acordo com o tipo de pele: se na pele clara os melanossomos se concentram em torno do centro do melanócito, na pele escura estes se distribuem ao longo de todo o queratinócito, maximizando a absorção da luz (Figura 4).

Enquanto que uma quantidade similar de feomelanina pode ser observada independente do tipo de pele, a proporção entre eumelanina e feomelanina é o que

decide a coloração dos cabelos (ITO; WAKAMATSU, 2011). Cabelos ruivos apresentam uma maior quantidade de feomelanina enquanto que em cabelos negros prevalece a eumelanina. Já cabelos loiros possuem um pouco de cada tipo de melanina (REES, 2003). No entanto, se os melanócitos presentes na pele possuem um ciclo de vida longo, aqueles presentes no folículo capilar permanecem por apenas 3-8 anos. Após esse tempo, os melanócitos no bulbo capilar são repostos a partir do estoque de células tronco presente (TOBIN, 2010).

Figura 4: Esquema da estrutura da pele humana: pigmentação clara (esquerda) e escura (direita). De cima para baixo: UV = radiação ultravioleta; SC = estrato córneo; G = estrato granuloso; S = estrato espinhoso; B = estrato basal; BM = membrana basal; D = derme. Tipos celulares: K = queratinócitos; M = melanócitos; F = fibroblastos; grânulos ovais = melanossomos contendo melanina. Fonte: (YAMAGUCHI; BRENNER; HEARING, 2007).

Em contraste ao que ocorre na pele e cabelos, nos olhos, os melanossomos não são secretados, ficando retidos nos melanócitos da íris (STURM; FRUDAKIS, 2004). Todas as colorações de olho apresentam uma quantidade semelhante de melanócitos. No entanto, o número de melanossomos presentes nesses melanócitos e por consequência, a quantidade de melanina produzida, varia de acordo com a coloração do olho (Figura 5). Além disso, a cor da íris se diferencia de acordo com as quantidades de eumelanina e feomelanina produzidas. Enquanto que mais eumelanina é encontrada em pessoas de olhos marrom, mais feomelanina é verificada em olhos azuis e verdes (KOLB, 2007).

Figura 5: Diferenças observadas de acordo com a coloração dos olhos. Destaque para os melanócitos e seus compartimentos (os melanossomos). Como pode ser observado, olhos azuis apresentam uma menor quantidade de melanossomos e maior quantidade de feomelanina, olhos verdes apresentam uma quantidade intermediária e olhos castanhos apresentam um maior número de melanossomos e maior quantidade de eumelanina (Adaptado de Sturm; Frudakis, 2004).

A rota bioquímica da melanogênese está sob complexo controle genético (STURM; DUFFY, 2012). Estima-se que mais de 120 genes estejam envolvidos neste processo, sendo possível observar uma grande variabilidade na pigmentação da pele entre as populações humanas e até mesmo dentro delas (PARRA; KITTLES; SHRIVER, 2004; PARRA, 2007; ROUZAUD et al., 2005; STURM, 2009). Este alto grau de variação pode ser atribuído primariamente ao efeito da Seleção Natural, que teria moldado a distribuição da pigmentação da pele nas populações humanas de acordo com um gradiente latitudinal de radiação ultravioleta (UVR) (PARRA, 2007). Segundo essa teoria, a cor da pele tende a ser mais escura em regiões equatoriais e tropicais devido à alta incidência de UVR, uma vez que a melanina atua como um protetor solar natural, protegendo o DNA e proteínas dos efeitos de ondas eletromagnéticas curtas (300nm) (PARRA, 2007). Em altas latitudes, a cor da pele tende a ser mais clara devido à menor incidência de radiação ultravioleta (UV), a qual é essencial para síntese de vitamina D, sendo a síntese via exposição solar a principal fonte deste recurso (HOLICK, 2005). Dessa forma, de maneira geral, a distribuição da pigmentação humana é resultado de um equilíbrio entre seleção natural favorecendo a proteção contra queimaduras em regiões onde a incidência de radiação UV é alta e seleção privilegiando

a pigmentação clara em regiões distantes do Equador, de modo a facilitar a síntese de vitamina D (HUDJASHOV; VILLEMS; KIVISILD, 2013).

Como consequência, diferentes populações apresentam diferentes *backgrounds* haplotípicos nos genes de pigmentação (CERQUEIRA et al., 2014). A população brasileira possui um *background* genético bastante heterogêneo, resultado de cinco séculos de miscigenação entre três principais grupos parentais: ameríndios nativos, africanos que foram trazidos como escravos e colonizadores europeus (LINS et al., 2010; MANTA et al., 2013; PENA et al., 2011).

1.3 DIVERSIDADE DA POPULAÇÃO BRASILEIRA

Com uma área total de 8.511,960 km² o Brasil é um país de dimensões continentais e possui diferentes regiões, caracterizadas por histórias populacionais distintas. Por exemplo, o Norte possui grande influência ameríndia, o Nordeste, uma forte presença africana e o Sul contou com um grande contingente europeu durante sua formação (PENA et al., 2011).

De acordo com o IBGE (Instituto Brasileiro de Geografia e Estatística) a distribuição dos grupos populacionais, com base na cor da pele autorelatada, é heterogênea entre as regiões geopolíticas brasileiras. Na classificação adotada desde 1991 pelo IBGE os indivíduos são agrupados em cinco categorias: “branco”, “pardo”, “negro”, “amarelo” e indígena. Os resultados do censo realizado em 2010 (<http://www.censo2010.ibge.gov.br/>) indicam que dos 190 milhões de brasileiros entrevistados, 47,7% foram classificados como “brancos”, 43,1% como “pardos”, 7,6% como “negros”, 1,1% como “amarelos” e 0,4% como indígenas. No estado de São Paulo, dos cerca de 40 milhões entrevistados, 63,9% se autodeclararam “brancos”, 29,1% “pardos”, 5,5% “negros”, 1,4% “amarelos” e 0,1% indígenas. Na cidade de Ribeirão Preto, que conta com pouco mais de 600 mil habitantes e constitui a principal fonte de origem de amostras do presente estudo, as frequências de “brancos”, “pardos”, “negros”, “amarelos” e indígenas foram respectivamente de 69,8%, 22,8%, 6,4%, 0,9% e 0,1% (<http://www.ibge.gov.br/cidadesat/link.php?codmun=354340>).

No entanto, de acordo com estudos de marcadores genéticos a variabilidade existente entre as diferentes regiões do Brasil é menor que aquela indicada pelos resultados do IBGE. Segundo estudo realizado por Pena et al. (2011) a ancestralidade europeia foi predominante em todas as regiões, com proporções que variaram de 60,6%

no Nordeste a 77,7% no Sul do país. Estes resultados indicam que a heterogeneidade da população brasileira não pode ser adequadamente representada através do critério arbitrário de categorias de cor, como sugerido pelo IBGE. Além disso, esta forma de classificação também pode variar ao longo do tempo e conforme o sexo, condição econômica e região a que pertence o classificador (LEITE et al., 2011; MANTA et al., 2013; SANTOS et al., 2009).

Por isso, embora correlacionadas, é preciso cautela ao utilizar a ancestralidade como preditor de pigmentação (e vice-versa), como ressaltado no estudo realizado por Shriver, et al. (2003), onde muitos marcadores se mostraram associados à pigmentação não por possuírem um efeito funcional, mas por serem informativos de ancestralidade. Neste caso, a associação entre tais marcadores e a pigmentação humana não seria extrapolável para populações autóctones ou que apresentem uma dinâmica distinta de miscigenação (SHRIVER et al., 2003).

1.4 MARCADORES GENÉTICOS NA PREDIÇÃO DO FENÓTIPO DE PIGMENTAÇÃO

Pigmentação e ancestralidade podem estar correlacionadas devido à presença de estratificação por miscigenação (PARRA; KITTTLES; SHRIVER, 2004). Todavia, populações com miscigenação recente são propícias para identificação de genes responsáveis pela expressão de determinados fenótipos, quando estes genes se manifestam de forma diferenciada nas populações ancestrais. Um grupo de indivíduos miscigenados com fenótipo semelhante possui um maior compartilhamento de regiões genômicas com a mesma origem ancestral (TANG et al., 2006).

Além disso, o Desequilíbrio de Ligação (DL) em populações miscigenadas é maior do que o observado em relação às populações parentais (NORDBORG; TAVARÉ, 2002). Assim, sequências genômicas – que incluem marcadores em DL apresentando grandes diferenças de frequências haplotípicas em comparação com as populações parentais – podem ser utilizadas para inferir ancestralidade, com vantagens em relação à inferência de ancestralidade por AIMs individuais (SANKARARAMAN et al., 2008). Haplótipos multi-alélicos (minihaps), definidos como pequenas regiões moleculares compostas por três ou mais SNPs, tem a capacidade de identificar maior quantidade de informação relacionada à ancestralidade que o mesmo número de SNPs conseguiria capturar individualmente (PAKSTIS et al., 2012).

O conceito inerente à análise de minihaps também pode ser estendido à determinação de fenótipos específicos, tais como a pigmentação, por meio de seleção de haplótipos pertencentes a regiões genômicas envolvidas em vias bioquímicas que conduzem ao fenótipo em questão. Neste contexto, tais haplótipos apresentariam um maior poder de resolução do que SNPs individuais.

A utilização do *Next Generation Sequencing* (NGS) oferece grandes possibilidades para identificação de minihaps, que podem ser utilizados em diversas aplicações forenses (PAKSTIS et al., 2012). Até a implementação do NGS, os haplótipos só podiam ser obtidos a consideráveis custos experimentais ou via inferência computacional da fase haplotípica, sendo que os algoritmos utilizados para este fim normalmente não obtinham sucesso na análise em larga escala de SNPs, principalmente quando provenientes de estudos de associação em escala genômica (*Genome-Wide Association Studies* - GWAS) (HALDORSSON; AGUIAR; ISTRAIL, 2010).

Estudos em larga escala, como GWAS e aquele desenvolvido pelo 1000 *Genomes Project*, representam excelentes oportunidades para pesquisa genômica funcional, permitindo estabelecer de forma mais efetiva a correlação entre genótipo e fenótipo (MARDIS, 2008; VIA; GIGNOUX; BURCHARD, 2010). Entretanto, o conhecimento relacionado à predição fenotípica continua bastante limitado, principalmente com relação a populações fora do eixo Europa - América do Norte (LIU; WEN; KAYSER, 2013).

Embora pouco tenha sido feito em termos de haplótipos, diversas variantes genéticas foram individualmente associadas a fenótipos de pigmentação através de GWAS. Estes estudos têm identificado SNPs tanto em genes reconhecidamente envolvidos com pigmentação, como *MC1R*, *OCA2*, *SLC45A2*, *SLC24A5*, *TYR*, *TYRP1*, *KITLG*, *ASIP* (LIU; WEN; KAYSER, 2013; ZHANG et al., 2013) quanto em genes até então não implicados com a característica, como no *locus IRF4* (HAN et al., 2008; LIU; WEN; KAYSER, 2013; SULEM et al., 2007).

1.5 O GENE *IRF4*

Situado próximo ao Complexo Maior de Histocompatibilidade (MHC), o gene *IRF4* (6p25-p23) pertence ao grupo de fatores de transcrição da família *Interferon Regulatory Factor* (IRF), sendo aparentemente expresso exclusivamente em células do sistema imune e em melanócitos (GUALCO, M, WEISS, ML, BACCHI, 2010).

A família IRF atua na resposta imune contra infecções virais (GUPTA et al., 1999). Sendo amplamente expressos, modulam não somente a resposta à interferon, como estão envolvidos no crescimento celular, transformação neoplásica, indução da apoptose e desenvolvimento da resposta imune mediada por células T (HARADA; TANIGUCHI; TANAKA, 1998).

Fatores de transcrição da família dos IRF atuam junto a outros fatores de transcrição para expressão de genes-alvo. É provável que estes fatores interajam diretamente com modificadores de histonas, controlando o meio ambiente da cromatina dos genes-alvo. Ao fazê-lo podem afetar não somente eventos prévios da iniciação da transcrição como ainda influenciar passos subsequentes, como alongamento e terminação (OZATO; TAILOR; KUBOTA, 2007).

Embora não dependa da indução por interferon, como outros genes dessa família, IRF4 se liga ao elemento de resposta ao estímulo por interferon (ISRE), presente na região promotora de genes que respondem à interferon (ESCALANTE; AGGARWAL, 1998). IRF4 é considerado um mediador crítico do desenvolvimento de células dendríticas, mieloides e linfoides (OZATO; TAILOR; KUBOTA, 2007; SHAFFER et al., 2009). Este gene tem sido relacionado à maturação das células B, com função crítica nos processos de mudança de classe por recombinação e geração de células plasmáticas a partir dos centros germinativos e de células B de memória (KLEIN et al., 2006; MITTRUCKER et al., 1989). No mieloma múltiplo, IRF4 funciona como um oncogene, sendo sua expressão essencial para o desenvolvimento das células cancerosas (SHAFFER et al., 2008). Em alguns casos de mieloma múltiplo e linfomas de células T, o gene pode ainda estar translocado em regiões regulatórias da cadeia-pesada de imunoglobulinas e em função disso, desregulado (SHAFFER et al., 2009).

A proteína codificada por *IRF4* compreende uma única cadeia polipeptídica, que pode assumir duas isoformas possíveis, compostas por 451 (isoforma 1) ou 450 (isoforma 2) aminoácidos (XU et al., 2012). Estima-se, com base na sequência RefSeq do gene IRF4 (NG_027728.1), que este compreenda 9 éxons e 8 íntrons, com um tamanho total de 19.705pb, sendo 126 pb correspondentes à região 5' *untranslated region* (5'UTR) e 3.845 pb à região 3' *untranslated region* (3'UTR).

Alelos e genótipos de polimorfismos em IRF4 foram associados com melanoma (DUFFY et al., 2010), leucemia linfocítica crônica (CROWTHER-SWANEPOEL et al., 2010) e ainda com condições não relacionadas ao sistema hematopoietico, como a

doença celíaca (DUBOIS et al., 2010). Poucos estudos tem associado variações neste *locus* a características de pigmentação (PRAETORIUS et al., 2013). Dos SNPs de *IRF4* relacionados à pigmentação, rs12203592 (C>T) é o que mostra a mais forte associação, sendo o alelo T associado com a presença de sardas, olhos azuis, cabelos castanhos e alta sensibilidade da pele à exposição solar em populações europeias (SULEM et al., 2007; HAN et al., 2008; DUFFY et al., 2010). Este SNP é polimórfico apenas na população europeia, apresentando frequências alélicas que variam de 3% na população finlandesa (FIN) até 18% na população da Inglaterra e Escócia (GBR). (http://browser.1000genomes.org/Homo_sapiens/Variation/Summary?r=6:395821-396821;source=HGMD-PUBLIC;v=CS098488;vdb=variation;vf=59170107) (DUFFY et al., 2010; WALSH, et al., 2011).

Tal variação rs12203592, localizada no íntron 4 do gene *IRF4*, influencia a função de um elemento cis-regulatório. O íntron 4 de *IRF4* possui um *enhancer* que regula a expressão deste gene em melanócitos. Em melanócitos, MITF e o fator de transcrição ativador da proteína alfa-2 (TFAP2 α) se ligam ao *enhancer* no íntron 4, aumentando a transcrição de *IRF4*. Em conjunto com MITF, o fator de transcrição *IRF4* se liga à região promotora de *TYR*, ativando sua transcrição (Figura 6).

Figura 6: Modelo da relação entre MITF, TFAP2 α e *IRF4* na expressão de *TYR* Fonte: Praetorius, et al., 2013.

Além dos sítios de ligação para MITF encontrados no íntron 4, o gene *IRF4* apresenta diversos outros potenciais sítios de ligação para MITF localizados cerca de 2kb *upstream* ao sítio de início da transcrição (Figura 7). Além disso, três sítios adicionais são encontrados no íntron 4 de *IRF4*. Por outro lado, não foram encontrados sítios de ligação para MITF na região promotora proximal do gene (STRUB, et al. 2011).

Figura 7: Análise dos potenciais sítios de ligação para fatores de transcrição no gene *IRF4*. O íntron 4 é mostrado em detalhe. TSS: sítio de início da transcrição. Fonte: Praetorius, C. The role of MITF in regulating human pigmentation, University of Iceland, Faculty of Medicine, 2014.

Como mencionado, foi identificado um sítio de ligação para TFAP2 α no local que inclui o SNP rs12203592 (Figura 7). O alelo rs12203592*T interrompe a ligação de TFAP2 α junto ao íntron 4 do gene *IRF4*, o que, por consequência, diminui a expressão de *IRF4*, afetando a ativação de TYR (PRAETORIUS et al., 2013). Este achado corroborou experimentos anteriores que já haviam demonstrado que TFAP2 α reconhece com menor afinidade o alelo T (DO et al., 2010). Uma vez que TYR cataliza os passos iniciais da biossíntese de melanina, a redução na sua expressão conduz a fenótipos de pigmentação mais claros. Deste modo, alterações em elementos regulatórios do DNA, como exemplificado pelo SNP rs12203592, são fatores que também podem influenciar na variação fenotípica.

Outros estudos indicam que a atividade dos IRFs também é regulada em momentos posteriores à etapa de transcrição, sendo que até mesmo modificações pós-traducionais foram relatadas (OZATO; TAILOR; KUBOTA, 2007). Dentre os mecanismos de controle pós-transcricional destaca-se o papel dos micro-RNAs (miRNAs) - moléculas de RNA não-codificante que se ligam por complementariedade de bases a sequências de mRNA, particularmente na região 3'UTR, diminuindo a

abundância ou a eficiência de tradução desse mRNA (BARTEL; LEE; FEINBAUM, 2004).

Estima-se que cerca de 30% dos genes codificadores em humanos sejam diretamente regulados por miRNAs (LEWIS; BURGE; BARTEL, 2005). Estas moléculas representam uma das classes mais bem caracterizadas de pequenos RNAs, embora ainda exista muitos aspectos a serem descobertos (DA CRUZ; JASIULIONIS, 2012). Até o momento, foram identificados mais de 2.588 dessas moléculas em humanos, release 21, junho 2014 (<http://www.mirbase.org/>).

Além disso, uma mesma molécula pode regular uma ampla variedade de fenótipos e vias metabólicas, indicando que a presença de um único miRNA é insuficiente para prever uma resposta celular (JAYANTHY; SETALURI, 2015). Por isso, a identificação das redes regulatórias que governam o efeito dos miRNAs na pigmentação humana se faz necessário. Iniciativas desta natureza, avaliando miRNAs que possuem afinidade à região 3'UTR de *IRF4* e tendo como alvo o processo de melanogênese, não foram realizadas até o momento.

2. JUSTIFICATIVA

O gene *IRF4* está envolvido na biossíntese de melanina (PRAETORIUS et al., 2013). Sendo seu produto um fator de transcrição e tendo em vista a influência desempenhada por elementos regulatórios na pigmentação humana é possível que esta molécula exerça funções ainda não estabelecidas na biologia do melanócito. Variações em suas regiões regulatórias podem afetar a expressão e disponibilidade deste fator de transcrição (por exemplo, influenciando a interação do MITF ou outros fatores de transcrição com a região regulatória de *IRF4*), enquanto que variações em sua região codificante podem afetar a estrutura da molécula *IRF4* e sua capacidade de ligação com as sequências regulatórias de genes alvos (*TYR*, por exemplo). Em ambos os casos, a regulação da expressão de genes alvos seria, conseqüentemente, alterada. No entanto, o conhecimento acerca da regulação transcricional em melanócitos e a forma como esses processos influenciam na pigmentação continua limitado (VISSER et al., 2014).

A utilização de tecnologias de genotipagem de alto-rendimento, como o NGS, combinadas a novas abordagens de bioinformática possui grande potencial em identificar novos sítios de variação relacionados a características de pigmentação humana (LIU; WEN; KAYSER, 2013), o que pode contribuir, por exemplo, para predição de características fenotípicas, úteis em aplicações forenses, para o entendimento da biologia do câncer de pele e na definição de novos marcadores de ancestralidade.

Considerando a grande heterogeneidade genotípica e fenotípica observada na população brasileira, o que aumenta o poder de detecção de novas variantes (SULEM et al., 2007), destaca-se que a utilização do NGS deverá aumentar o conhecimento acerca da diversidade genética e, por consequência, da função de *IRF4* nos melanócitos humanos.

3. HIPÓTESE

Variações presentes nas regiões regulatórias (promotora, 5'UTR e 3'UTR) e codificadora de *IRF4* ou haplótipos envolvendo essas variações são correlacionados a efeitos na pigmentação humana, uma vez que interações com os genes que regulam a expressão de IRF4 (MITF, TFAP2 α , por exemplo) ou que são regulados por tal molécula (TYR) seriam afetadas.

4. OBJETIVOS

Este trabalho tem como objetivo principal avaliar, através do sequenciamento de nova geração, a diversidade genética das regiões promotora, 5'UTR, codificadora e 3'UTR do gene *IRF4* em amostras de 228 indivíduos da população do Estado de São Paulo. Com isso, será possível:

- Identificar novos polimorfismos do gene *IRF4*;
- Caracterizar a região regulatória (promotora, 5'UTR e 3'UTR) e codificadora (éxons e parte dos íntrons) do gene *IRF4*;
- Determinar as frequências alélicas, genotípicas e a presença de haplótipos compostos por polimorfismos do gene *IRF4*;
- Avaliar a intensidade do Desequilíbrio de Ligação (LD) entre a região regulatória e codificadora do gene *IRF4*;
- Comparar a variabilidade do gene *IRF4* entre grupos resultantes da estratificação das amostras de acordo com a pigmentação da pele, olhos, cabelos e presença de sardas;
- Identificar um conjunto de polimorfismos do gene *IRF4* que contribuam para determinação da pigmentação da pele, olhos, cabelos e presença de sardas, em indivíduos da população brasileira;
- Avaliar a relação entre as medidas objetivas dos índices de melanina de cada indivíduo com os SNPs associados a alguma das características de pigmentação avaliadas.

5. MATERIAL E MÉTODOS

5. MATERIAL E MÉTODOS

5.1 COLETA DAS AMOSTRAS

Conforme previsto pelo Projeto aprovado junto ao Comitê de Ética em Pesquisa da USP (Processo CEP – FFCLRP – CAAE n.25696413.7.0000.5407), o termo de Consentimento Livre e Esclarecido (TCLE, ANEXOS 1 e 2) foi obtido de todos os 228 participantes do estudo, recrutados junto ao Hemocentro de Ribeirão Preto – SP.

Cada participante respondeu a um questionário, de onde foram obtidos os dados relacionados a fenótipo e ancestralidade e doou uma amostra de material biológico, que consistiu em 10 mL de sangue periférico, mantido em tubo *Vacutainer*® com anticoagulante EDTA.

O grupo de participantes, composto por indivíduos não aparentados, de ambos os sexos, com idade variando entre 18 e 64 anos e representando os grupos ancestrais da população brasileira, foram classificados em categorias de acordo com a cor da pele, cor dos olhos (azul, verde, mel, castanho claro e castanho escuro), cor natural do cabelo (ruivo, loiro-claro, loiro-escuro, castanho-claro, castanho-escuro e preto), porcentagem de cabelos grisalhos (ausência, 0-25%, 25-50%, 50-75% e 75-100%) e quanto a presença de sardas (efélides). A classificação dos indivíduos com relação à pigmentação da pele foi realizada conforme o sistema proposto por Fitzpatrick (1988), onde a pontuação mais baixa (I) representa pele muito clara e sensível à UVR e a maior pontuação (VI) representa pele escura, que nunca se bronzeia em reação à UVR. A classificação da cor da pele, olhos e cabelos foi feita por cada participante (autoclassificação) e ainda de forma independente por dois membros da pesquisa (heteroclassificação), chegando-se ao final a um consenso quanto à heteroclassificação dos indivíduos. Informações pessoais associadas aos dados pessoais de cada indivíduo foram organizadas em planilhas formato.xls.

Adicionalmente, a intensidade de pigmentação da pele (testa e região interna do braço) foi quantificada de forma objetiva por meio do índice de melanina obtido a partir do espectrofotômetro CM2500D e do software *SkinAnalysis* (*Konica Minolta Holdings, Inc*), aumentando a acurácia da determinação do fenótipo. Imagens digitais dos olhos e cabelos foram obtidas por uma câmera fotográfica *Canon SX510HS*. Abertura, *shutterspeed* e ISO foram fixados para $f = 3,4$, $1/10$ e 100, respectivamente, tendo sido utilizada uma distância média de 7 cm entre a câmera e o indivíduo para obtenção das imagens fotográficas.

5.2 ANÁLISE LABORATORIAL

5.2.1 Extração do DNA genômico

Após a coleta, a extração do DNA foi 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 foi transferido para um tubo *falcon* graduado de 50mL, ao qual foi 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 final. De forma a lisar as hemácias, o conteúdo foi delicadamente misturado por imersão e centrifugado durante 6 minutos (min) a 3300 rotações por minuto (rpm) em centrífuga Excelsa™ II modelo 206BL, FANEM™, São Paulo, Brasil. A extração prosseguiu com o botão de células obtido, o qual foi ressuspensionado em 4,5mL de tampão de lise II (NaCl 0,075M; NaEDTA 0,024M), 125μL de SDS 10% e 1,1mL de perclorato de sódio (5M) e vortexado a temperatura ambiente durante 10min para rompimento dos glóbulos brancos. Para a lise das proteínas foram adicionados 2mL de NaCl 6M ao tubo *falcon*, que foi vortexado a temperatura ambiente por 15 segundos. A seguir, o conteúdo foi centrifugado por 8min a 3500 rpm. O sobrenadante foi transferido para um novo tubo de 50mL, ao qual foi adicionado 5mL de isopropanol absoluto (Merck Millipore, Darmstadt, Alemanha) e homogeneizado até que o DNA precipitasse.

Finalmente, o DNA precipitado foi 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 foi então centrifugado por 5min a 5000 rpm e o sobrenadante descartado a seguir. O microtubo foi deixado em centrífuga a vácuo (Thermo Savant™ DNA120) durante 10min para remoção completa do etanol. Por fim, o DNA aderido ao microtubo foi dissolvido em 200μL de água MilliQ, mantido na geladeira por 1 semana e posteriormente armazenado a -20°C no banco de amostras do laboratório.

5.2.2 Preparo dos reagentes para enriquecimento de alvo

Regiões-alvo, incluindo a sequência promotora, CDS (*Coding DNA Sequence*), 5'UTR e 3'UTR do gene *IRF4* e de outros genes relacionados à pigmentação humana

(*MC1R*, *OCA2*, *HERC2*, *SLC45A2*, *SLC24A5*, *TYR*, *TYRP1*, *DCT*, *MITF* e *ASIP*) foram submetidas ao ensaio *in silico* de desenho de sondas para captura de DNA alvo utilizando o aplicativo *SureDesign* da *Agilent* (<https://earray.chem.agilent.com/suredesign/home.htm>). Para desenho das sondas, o *SureDesign* toma 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, as sondas desenhadas capturam uma região de 488.658pb, o que garante o sequenciamento parcial dos introns flanqueadores dos exons alvo.

O *NanoDrop® ND-1000* (*Thermo Fisher Scientific Inc.*) foi 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, foi 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 foram quantificadas através de fluorescência (*Qubit - Life Technologies*) por meio do *Qubit™ dsDNA BR Assay*. Para isso, o DNA genômico foi diluído a uma proporção de 1:10 (6µL DNA: 54µL de água Milli-Q). Inicialmente, foi preciso fornecer ao instrumento uma curva padrão. Para tanto, o corante de dsDNA foi diluído no buffer do Qubit (1:200) e 10µL dos padrões (1 e 2) foram diluídos em 200µL da solução anterior. Realizada a leitura, uma curva padrão foi estabelecida, a partir da qual foi calculada a concentração da amostra original, com base no volume adicionado para leitura. Todas as amostras foram diluídas à concentração de 5ng/µL, no intuito de normalizar as amostras a uma concentração compatível para a montagem das bibliotecas de fragmentos.

5.2.3 Enriquecimento do alvo e preparo da biblioteca

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

fossem 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.

5.2.3.1 Digestão e hibridização do DNA

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

Para o mix de restrição, preparado para lotes de 12 amostras (11 amostras de DNA genômico e outra de ECD), foram adicionados 476µL de *RE buffer*, 11,9µL de BSA e 7µL de cada um dos dois conjuntos com 8 enzimas de restrição. Deste mix, foi distribuído 35µL em cada tubo contendo a amostra de DNA genômico, totalizando um volume final de 80µL. A mistura contendo o mix de restrição e o DNA genômico foi então incubada no termociclador (*SureCycler 8800 Thermal Cycler, Agilent Technologies*) a 37°C por 30 minutos.

Para o mix da fase de hibridização foram adicionados 650µL da *Hybridization Solution* e 260µL das sondas *Haloplex*, sendo distribuídos 70µL dessa solução por amostra. Adicionamos 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).

O DNA digerido de cada amostra (80µL) foi então transferido para os microtubos contendo o mix de hibridização (70µL) e seus índices correspondentes (10µL). Para a fase de hibridização das sondas biotiniladas, essas amostras foram incubadas a 54°C durante 16 horas. O tempo de duração da hibridização é determinado de acordo com a densidade de sondas do ensaio. Segundo instruções do fabricante (*Haloplex, Agilent Technologies*), ensaios que atingem cobertura de entre 500 kb - 5 Mb exigem 16 horas de hibridação, enquanto que ensaios com cobertura de entre 1-500 kb exigem apenas 3 horas de hibridação. Apesar de o presente ensaio estar voltado para a

cobertura de apenas 488,7 kb, o elevado número de sondas desenhadas determinou o tempo de hibridização de 16 horas.

5.2.3.2 Captura, ligação, PCR e eluição

Para a fase de captura, 520µL de beads magnéticas recobertas por estreptavidina foram adicionadas a um tubo de 1,5 mL, o qual foi acondicionado junto a uma placa magnética (*DynaMag™-2 Magnet, Life Technologies*), sendo o sobrenadante da solução removido a seguir. Junto às *beads* concentradas adicionamos 520µL de *Capture Solution*. Deste conteúdo, 40µL foi adicionado a cada solução de DNA hibridizado (160µL) e após 15min, os tubos contendo as soluções foram colocados junto a uma placa magnética de 96 poços (*Agencourt SPRIPlate 96R, Beckman Coulter*). A seguir, adicionamos 100µL de *Wash Solution* por amostra, sendo a solução incubada no termociclador a 46°C por 10min. Após esse período, o excesso foi novamente removido.

Para a fase de ligação adicionamos 617,5 µL de *Ligation Solution* e 32,5µL de DNA ligase e distribuímos 50 µL dessa solução a cada amostra hibridizada e capturada. A solução foi então incubada a 55°C por 10min para preenchimento dos *gaps* e ligação.

A seguir, eluímos o conteúdo de DNA ligado com 100 µL de *SSC buffer*, retirando o sobrenadante ao final da etapa com o auxílio da placa magnética. Para eluir o DNA das *beads* adicionamos 25 µL de NaOH (50 mM). Com o auxílio da placa magnética, coletamos o sobrenadante num total de 20 µL por amostra e o transferimos para o mix de PCR. Tal mix para a reação de PCR foi preparado com 209,3 µL de água MilliQ, 130 µL de *5X Herculanase II Reaction*, 5,2 µL de dNTPs (100 mM, 25 mM para cada dNTP), 13 µL de cada um dos *primers forward* e *reverse* (25 µM), 6,5 µL de ácido acético (2M) e 13 µL da polimerase *Herculanase II Fusion*, para um total de 30 µL por amostra. O volume final da reação foi de 50 µL por amostra. As condições de ciclagem da PCR seguiram uma etapa inicial de desnaturação por 10min a 98 °C, seguida por 20 ciclos de 98 °C por 30seg, 60 °C por 30seg e 72 °C por 1min, e um passo final a 72 °C por 10min e 8 °C infinito.

5.2.3.3 Purificação da biblioteca

Os produtos de PCR foram então purificados utilizando *AMPure XP beads* (*Beckman Coulter*) de acordo com as instruções da *Agilent*. Um total de 40 µL do

produto de PCR foi misturado a 100 μL de *AMPure XP beads* mais 40 μL de água MilliQ por amostra. A seguir, essa solução foi colocada junto à placa magnética para remoção do sobrenadante e foi lavada duas vezes com álcool 70%, que foi completamente removido ao final. Então adicionamos 40 μL de Tris-HCl (pH = 8,0) para eluir o DNA capturado pelas *AMPure XP beads*. O produto final de 40 μL foi armazenado a -20°C .

5.2.4 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 da sequência (CHER, 2011).

O controle de qualidade das bibliotecas foi 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 8). A partir da seleção manual da faixa de fragmentos gerados (175 a 625pb) é possível obter a concentração (pg/ μL) e a molaridade (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 sequenciamento.

O *2100 Bioanalyzer* foi utilizado também para quantificação do *pool* final de bibliotecas sequenciadas no primeiro e segundo lotes (48 e 95 amostras, respectivamente). O *pool* final de bibliotecas sequenciadas no terceiro lote (87 amostras) foi quantificado tanto pelo *2100 Bioanalyzer* quanto pelo *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).

Figura 8: Perfil de picos esperado no eletroferograma para as bibliotecas de fragmento deste ensaio. Tal eletroferograma foi 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 (35 pb) e do maior (10380 pb) marcador apresentem um tempo de corrida correspondente entre as amostras analisadas e o *ladder* e que as linhas de base sejam planas (Cher, C. *Bioanalyzer Applications for Next-Gen. Sequencing: Updates and Tips*, *Agilent Technologies*, 2011, 49p).

5.2.5 Sequenciamento

Como descrito no item anterior, a partir das concentrações obtidas no *2100 Bioanalyzer* (primeiro e segundo lotes) e *Qubit* (terceiro lote), preparamos soluções a 4 nmol/L para cada biblioteca. 10 µL de cada amostra foram misturados em um único microtubo, gerando o *pool* de bibliotecas, o qual foi quantificado utilizando-se o *2100 Bioanalyzer* (primeiro e segundo lotes) e qPCR (terceiro lote). A seguir, 5 µL do *pool* de bibliotecas a 4 nmol/L foram desnaturados pela mistura com 5 µL de NaOH (0,2M) e 990 µL de *Hybridization Buffer* (HT1) para obtenção de uma biblioteca a 20 pM, que foi diluída (a partir do volume de 480 µL) com 120 µL de HT1 para que se chegasse à concentração final de 16 pM. Tal material foi inserido 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.

Inicialmente as moléculas de DNA fita simples são ligadas em uma região específica da superfície da *flow cell* de acordo com seus adaptadores. 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 imagem de 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 identifica de maneira unívoca o fragmento a uma determinada amostra.

5.3 ANÁLISE DE DADOS

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 do consenso a um genoma de referência.

5.3.1 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, 1998), se relaciona à probabilidade de erro em uma determinação de 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 registro do arquivo FastQ contém quatro linhas:

1. O identificador iniciado com um sinal @. Esta identificação segue o formato:
 - @Instrumento:IDCorrida:IDFlowCell:Linha:Canaleta:X:Y
ReadNum:FilterFlag:0:NúmeroAmostra
2. A sequência
3. Um sinal positivo (+)
4. Os valores de qualidade codificados no formato ASCII

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<???#
```

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, e o último segmento 001 sempre recebe essa denominação:

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

Como pode ser observado, os arquivos FastQ são obtidos de forma compactada, como indicado pelo (*.gz) no nome do arquivo. Para que possa ser dada continuidade às análises dos dados, é necessário a descompactação do arquivo, que pode ser obtido através da linha de comando a seguir em sistema operacional *Linux*:

```
gunzip PIG0001_S1_L001_R1_001.fastq.gz > PIG0001_S1_L001_R1_001.fastq
gunzip PIG0001_S1_L001_R2_001.fastq.gz > PIG0001_S1_L001_R2_001.fastq
```

5.3.2 Análise secundária

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

5.3.2.1 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, foi empregado o *software* FastQC (ANDREWS, 2010), o qual utiliza um arquivo FastQ para gerar um relatório completo da qualidade das sequências (Figura 9a). A partir dos resultados apresentados é possível decidir por estratégias que melhorem a qualidade dos dados gerados.

Figura 9: Gráfico de distribuição da qualidade por base no programa FastQC. Cada coluna apresenta um boxplot com os valores de qualidade de todas as *reads* em uma posição indicada. 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*.

5.3.2.2 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 bases 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* 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, 301 ciclos). Mesmo que um *cluster* inclua um pequeno fragmento a ser sequenciado, os ciclos de sequenciamento não são interrompidos, continuando até alcançar as sequências dos adaptadores. Como consequência, as *reads* geradas como *output* contêm a sequência da molécula de interesse e ainda a sequência do adaptador. 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.

No ensaio utilizado foram empregados adaptadores do tipo *paired-end* e uma sequência *barcode* por amostra, o que permitiu o sequenciamento simultâneo de várias amostras em uma única linha da *flow cell*. Como os *barcodes* são sequenciados em oito ciclos após o sequenciamento dos *reads forward* e antes do sequenciamento dos *reads reverse*, é possível designar de maneira unívoca cada par de *reads* à amostra correta. A sequência dos adaptadores do ensaio *Haloplex* para sequenciamento *Illumina* é conhecida, permitindo que estes sejam facilmente removidos da sequência.

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 na sequência, neste caso os *primers* da PCR e o *barcode* quando presente.

A trimagem pelos adaptadores foi 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 > logfile
```

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.
> logfile	permite redirecionar os registros de saída dos arquivos <i>output</i> para um único arquivo log.

5.3.2.3 Alinhamento ao genoma

Estando as *reads* processadas, o próximo passo foi 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 foi utilizado o BWA (LI et al., 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: BWA-backtrack, BWA-SW e BWA-MEM.

Antes de realizar o alinhamento, porém, é preciso indexar o genoma de referência por meio do índice FM (FERRAGINA et al., 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.

O algoritmo utilizado para realizar o alinhamento foi 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 foi utilizada (Tabela 2):

```
bwa mem -t 6 -R
' \@R\G\tID:sampleID\tLB:sampleLB\tSM:tag\tPL:illumina\tPU:samp
leP' /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.
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.

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

Comandos utilizados no BWA	Função do comando executado
/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 foi convertido no formato BAM (*.bam).

5.3.2.3.1 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 e trazem como informações principais (Tabela 3):

Tabela 3: Descrição dos componentes de um arquivo SAM

Seções que compõem um arquivo SAM	Descrição
HD	Cabeçalho
SQ	Dicionário da sequência de referência: traz informações sobre a referência, como nome e tamanho da sequência
RG	Grupo da <i>read</i> : traz informações sobre a corrida de sequenciamento, como linha ocupada pela amostra no sequenciamento, amostra, biblioteca)
PG	Programa utilizado para o alinhamento: contém informações como nome do alinhador, versão e parâmetros utilizados para o alinhamento

Cada *read* alinhada à referência é representada em uma linha da seção de alinhamento. Cada linha contém 11 seções obrigatórias, que devem sempre aparecer na ordem indicada na Tabela 4.

Tabela 4: Campos obrigatórios de uma linha da seção de alinhamento de um arquivo SAM. Fonte: *The SAM/BAM Format Specification Working Group, Sequence Alignment/Map Format Specification*, 2015, <http://samtools.github.io/hts-specs/SAMv1.pdf>

Col	Área	Descrição
1	QNAME	Nome do molde interrogado
2	FLAG	Marcação <i>bitwise</i>
3	RNAME	Nome da sequência de referência
4	POS	Posição de mapeamento (-1 base) à esquerda
5	MAPQ	Qualidade de mapeamento
6	CIGAR	CIGAR <i>string</i>
7	RNEXT	Nome do <i>mate</i> ou <i>read</i> seguinte
8	PNEXT	Posição do <i>mate</i> ou <i>read</i> seguinte
9	TLEN	Tamanho do molde
10	SEQ	Sequência da <i>read</i>
11	QUAL	Valores de qualidade Phred codificados no formato ASCII

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”).

5.3.2.3.2 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, foi preciso ordenar os alinhamentos de acordo com a posição

genômica. O comando a seguir foi 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 FIG0001 representa o nome da amostra e *sorted* indica que o arquivo BAM gerado se encontra ordenado:

```
FIG0001.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 foi nomeado como a seguir:

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

Uma vez ordenados e indexados, os arquivos BAM puderam ser visualizados através de um visualizador externo. Neste trabalho utilizamos o *Integrative Genomics Viewer* (IGV) 2.3.67, um programa em java disponibilizado pelo *Broad Institute* (ROBINSON et al., 2011; THORVALDSDOTTIR; JAMES; JILL, 2012). Um exemplo dessa visualização é apresentado na Figura 10.

Figura 10: Representação dos alinhamentos das *reads* na região do gene *IRF4* no programa IGV_2.3.67. A parte principal da tela de visualização é dividida em três partes. No topo pode ser visto representação esquemática do cromossomo 6 e a região deste cromossomo que está sendo visualizada (indicada por uma barra vermelha). Logo abaixo, é apresentada uma escala contendo as coordenadas da região. Na parte intermediária da tela é mostrada parte do arquivo BAM carregado. A visualização desse arquivo é dividida em duas regiões; acima, temos um gráfico representativo da cobertura (ou número de *reads* cobrindo cada base do genoma) e abaixo, os alinhamentos das *reads*. Na parte inferior da tela temos a anotação do genoma para a região visualizada (*RefSeq Genes*).

5.3.3 Análise terciária

5.3.3.1 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* (MCKENNA et al., 2010).

O programa apresenta uma estrutura modular e é dividido em diversas análises, as quais são reunidas junto a um *framework* (captura a funcionalidade comum a várias aplicações) 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 foi utilizada a ferramenta *HaplotypeCaller* do GATK. Considerando simultaneamente todas as amostras a serem analisadas, 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 haplótipo contra o haplótipo de referência com o objetivo de identificar potenciais sítios de variação.

Comparado ao *UnifiedGenotyper*, outra ferramenta que realiza a determinação de genótipos utilizando modelos separados para cada tipo de variação, 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. 2014).

A seguinte linha de comando foi utilizada para realizar a nominação das variantes (Tabela 5):

```
java -jar GenomeAnalysisTK.jar \
  -R reference.fasta \
  -T HaplotypeCaller \
  -I listofsamples.list \
  --dbSNP dbSNP.vcf \
  -stand_call_conf 30 \
  -stand_emit_conf 10 \
  -L targets.interval_list \
  -o output.raw.snps.indels.vcf \
  --dontUseSoftClippedBases \
  -drf DuplicateRead
```

O *HaplotypeCaller* gera como *output* um arquivo genômico intermediário (gVCF), contendo todos os sítios avaliados como variáveis. Além de permitir a identificação dos genótipos, a metodologia utilizada tornou possível a obtenção de alguma informação sobre a fase dos haplótipos, que identifica os alelos que estão localizados em um mesmo cromossomo.

Tabela 5: 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. Neste caso, o <i>HaplotypeCaller</i>
-I	indica arquivo de texto com lista de amostras (formato BAM) a ser utilizado pelo <i>HaplotypeCaller</i>
--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
-stand_emit_conf	permite emitir determinações de genótipo com baixa qualidade como resultados filtrados
-L	direciona o GATK a restringir o processamento dos dados aos intervalos genômicos especificados no arquivo “ <i>targets.interval_list</i> ”, que neste caso engloba os intervalos a seguir: chr6 388587 - 389344 chr6 389475 - 390745 chr6 390923 - 390971 chr6 391601 - 391974 chr6 393013 - 393376 chr6 394678 - 395077 chr6 395757 - 396422 chr6 397087 - 397346 chr6 397363 - 397643 chr6 398675 - 399103 chr6 401389 - 401867 chr6 404957 - 405207 chr6 406640 - 406932 chr6 407371 - 411488
-o	indica o <i>output</i> , que é um arquivo VCF (<i>Variant Call Format</i>) bruto (não filtrado) contendo todos os sítios que o <i>HaplotypeCaller</i> 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).

5.3.3.2 Remoção de genótipos de baixa qualidade

Alguns dos genótipos considerados pelo *HaplotypeCaller* 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 foram interrogados, tomando como *input* um arquivo VCF determinado pelo *HaplotypeCaller*. Neste processo, o arquivo VCF foi tratado por uma ferramenta baseada em Perl, Vcfx (http://www.castelli lab.net/apps/apps_vcfx.php), que utiliza como regras:

- para que um homozigoto seja inferido é exigida uma cobertura mínima de sete *reads*; caso contrário, um *missing allele* é introduzido nesta posição. Este procedimento visa garantir que um genótipo identificado como homozigoto seja resultado de variabilidade inexistente naquela posição e não devido a um segundo alelo eventualmente não-amostrado;

- sendo um dos alelos representado em menos de 5% das *reads*, este genótipo é estabelecido como homozigoto para o alelo mais frequente. Este procedimento, aplicado apenas para *reads* com cobertura mínima de 20 vezes em relação à posição avaliada, visa reduzir o nível de erros de amplificação e sequenciamento;

- em genótipos em que um dos alelos está subrepresentado (entre 5-20%) foi introduzido um *missing allele* no lugar do alelo menos frequente. Este procedimento é válido em situações de baixa cobertura ($x \leq 20$), onde a presença de uma única *read* poderia indicar erro de amplificação ou sequenciamento (possível falso-positivo) ao invés da ocorrência de um alelo alternativo verdadeiro. Assim, a definição deste genótipo, seja como homo ou heterozigoto, será realizada por inferência durante um processo posterior de imputação;

- apenas posições em que a proporção do alelo menos frequente foi superior a 20% foram genotipadas como heterozigotos, assegurando que somente genótipos heterozigotos de alta qualidade passem adiante, para o processo de imputação.

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

```
vcfx uncertain -i input=input.vcf output=output.vcf
```

5.3.3.3 Faseamento experimental

Em seguida, foi aplicado o algoritmo *ReadBackedPhasing*, também incorporado no programa GATK, o qual se utiliza da informação de fase das *reads paired-end* para encontrar o haplótipo local com a mais alta probabilidade.

A linha de comando a seguir foi utilizada para realizar a identificação dos haplótipos (Tabela 6):

```
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 6: 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 gerado pelo VCFx
-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

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* foi incorporado ao

método de faseamento computacional PHASE 2.1 (STEPHENS; SMITH; DONNELLY, 2001).

5.3.3.4 VCF

O VCF é o formato de arquivo de escolha do GATK, que contém informações sobre variantes encontradas em posições específicas em relação a um genoma de referência.

O nome do arquivo segue o formato a seguir, onde # identifica o número da amostra, determinado pela ordem na qual as amostras são listadas na *SampleSheet*:

SampleName_S#.vcf

O arquivo VCF possui um cabeçalho, bastante útil, uma vez que contém uma descrição para os campos contidos no arquivo (Tabela 7):

Tabela 7: 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 foram 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 foi utilizado como <i>input</i> para gerar o arquivo BAM

Além do cabeçalho, um arquivo VCF contém as linhas com os dados, 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 são as seguintes:

Tabela 8: 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
AN	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

Continuação Tabela 8: Descrição dos componentes do corpo de um arquivo VCF

Informações presentes no corpo do arquivo	Descrição das abreviações apresentadas
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>

5.4 ANÁLISES ESTATÍSTICAS

5.4.1 Faseamento computacional

A partir das diversas configurações de haplótipo possíveis para os genótipos identificados foram estimadas as probabilidades de uma determinada configuração haplotípica. Dentre as abordagens estatísticas disponíveis para o faseamento de indivíduos não aparentados optou-se pela utilização do programa PHASE 2.1 (STEPHENS; SMITH; DONNELLY, 2001).

O PHASE implementa um método de estatística Bayesiana para reconstruir os haplótipos a partir dos dados de genotipagem da população (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. Foram utilizados os seguintes parâmetros para as análises realizadas no PHASE: número de interações (1000), *thinning interval* (1) e *burn-in value* (100).

Uma primeira corrida foi realizada utilizando os dados filtrados pelo VCFx (correspondendo a 107 variações identificadas). Adicionalmente, foram desenvolvidos *scripts* (E.C. Castelli) 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 (no presente caso, até 8 blocos). Em seguida, o programa PHASE foi rodado 7 vezes, sendo cada uma das seis primeiras levando em consideração a informação de um dos blocos e a sétima não levando em consideração quaisquer informações do faseamento experimental. Como muitos sítios de variação se mostraram não-polimórficos na presente amostra, levando o PHASE a inferir haplótipos com valores muito baixos de probabilidade, foi realizada uma nova análise excluindo-se os 31 *singletons* (contagem alélica =1) e outros 2 pontos de variação da região promotora distal, que apresentaram uma baixa qualidade de leitura. Com base nos resultados de haplótipos obtidos a partir da segunda análise do PHASE foi realizada a maturação visual dos resultados, sendo selecionados indivíduos ($n=170$) que tiveram o mesmo par de haplótipos inferido com probabilidade de pelo menos 60% em todas as 7 corridas do programa, resultando em uma probabilidade média de acurácia de 88,96%.

Feita essa seleção, foi realizada uma terceira corrida no PHASE especificando a fase completa destes 170 indivíduos, e permitindo com isso que o programa melhorasse seu desempenho na inferência dos haplótipos dos 58 indivíduos restantes. Para tanto, foi editado o arquivo contendo as fases conhecidas, onde as informações de fase daqueles indivíduos selecionados anteriormente foram substituídas por “0” (o que indica que a fase para todos os *loci* de determinado indivíduo é a mesma que a dos genótipos no arquivo *input*) e as informações de fase obtidas pelo *ReadBackedPhasing* mantidas naqueles 58 indivíduos para os quais gostaríamos que fosse inferida a fase dos haplótipos.

O arquivo *out.pairs* editado (contendo somente o par de haplótipos inferido com maior probabilidade para cada indivíduo) foi utilizado como *input* para rodar o *script* *phase_to_vcf.pl* (E.C. Castelli), que gerou como *output* um arquivo VCF faseado. Neste momento, as linhas com os *singletons* que tiveram a fase informada pelo *ReadBackedPhasing* foram reinseridas no arquivo VCF faseado.

5.4.2 Desequilíbrio de ligação

O arquivo VCF foi convertido para o formato Arlequin através do script `vcf_to_arlequin_global.pl` (E.C. Castelli).

O desequilíbrio de ligação entre os *loci* estudados foi 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 foi utilizado o programa *Haploview* 4.2 (BARRETT et al., 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 $D' \geq 0,98$ (o que é consistente com ausência de recombinação histórica) e o limite inferior do intervalo for $> 0,7$. Assim, um bloco haplotípico é definido como uma região onde uma pequena parcela das comparações entre os pares de SNP ($< 5\%$) 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 et al., 2002).

5.4.3 Frequências alélicas, equilíbrio de Hardy-Weinberg e heterozigose esperada

As frequências alélicas para cada SNP (tanto estratificadas por categoria fenotípica quanto globalmente) e a aderência das frequências genotípicas ao teste de Equilíbrio de Hardy-Weinberg (EHW) foram verificadas através do programa GENEPOP v.4.0.10 (ROUSSET, 2008). A heterozigose esperada foi calculada a partir do programa Arlequin 3.5 (EXCOFFIER; LISCHER, 2010).

5.4.4 Associações Genótipo-Fenótipo

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

A mesma planilha foi utilizada ainda para o cálculo de *Odds Ratio* (OR) e do intervalo de confiança associado (IC) (LE, 2003; van BELLE, et al., 2004). Foi adotado o IC de 95% e o limite de significância foi estabelecido em $\alpha = 0,05$. O valor de OR permitiu verificar quantas vezes o caráter em questão (fenótipo avaliado) foi mais frequente no grupo de indivíduos apresentando determinado genótipo/alelo/haplótipo, comparativamente ao risco em indivíduos não portador deste fator (RUMEL, 1986).

5.5 ANÁLISES DE MICRORNAS

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 (MARINN; VANNÍCEK, 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 *IRF4*. A restrição da análise à 3'UTR se deve ao fato de a maioria dos miRNAs se ligarem preferencialmente nessa região (LEWIS et al., 2003).

A fim de identificar miRNAs que se liguem a sequências polimórficas da 3'UTR de *IRF4* foram selecionados os haplótipos encontrados com frequência acima de 1% na nossa análise. Deste modo, selecionamos 13 haplótipos, que foram nomeados como UTR-1 a UTR-13 em função da ordem decrescente das frequências haplotípicas observadas.

Tendo em vista o grande número de falso-positivos obtidos através dos métodos *in silico* foram utilizados três algoritmos distintos para predição de miRNAs com afinidade a sequências da 3'UTR de *IRF4*: RNAhybrid (GIEGERICH et al., 2004), miRanda (JOHN et al., 2004) e intaRNA (BUSCH et al., 2008).

Considerando que o miRNA se hibridiza 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 foi empregado um pacote de *scripts* em *Perl*, miRP (<http://www.castelli-lab.net/index.php#>), que automatiza a análise como brevemente descrito: a) as sequências de mRNA da 3'UTR foram fragmentadas em uma série de subsequências de 500 nucleotídeos com uma sobreposição de 50 nucleotídeos iniciais e finais; b) os três algoritmos foram utilizados para prever interações miRNA/mRNA considerando cada uma das subsequências; c) os resultados gerados apresentaram todas as interações miRNA/mRNA encontradas por cada algoritmo; d) manteve-se nos resultados somente a ligação mais estável nos casos em que foi identificada 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* foram utilizados como referência, com uma tolerância máxima de 3 nucleotídeos); e) os resultados encontrados pelos três algoritmos foram comparados; f) um sistema de pontuação foi aplicado para cada interação miRNA/alvo. Os parâmetros utilizados para o miRanda foram: -4 para extensão e -9 para abertura do *gap*, limiar de pontuação foi estabelecido como 50 e a mínima energia livre foi configurada para -1 kcal/mol. Para o RNAhybrid: a mínima energia livre de ligação foi configurada para -1 kcal/mol e o número de *hits* foi ajustado para 10. O intaRNA é um método livre de parâmetro, então a configuração padrão foi utilizada (PORTO et al., 2015).

O banco de dados contendo os miRNAs humanos até então identificados (2.588) foi obtido a partir do miRBase, release 21 de junho de 2014 (GRIFFITHS-JONES et al., 2006).

Os algoritmos foram 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 foi aplicado para cada conjunto de dados,

como descrito detalhadamente em Porto et al. (2015). Essas pontuações tiveram 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 foi utilizado como referência do algoritmo para propósito de normalização. Este procedimento teve como objetivo tornar os MFEs dos diferentes algoritmos comparáveis. A pontuação de ϵ para cada interação miRNA/mRNA foi 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 foi obtida pela soma das pontuações de cada algoritmo: $\epsilon = \epsilon_{\text{RNAhybrid}} + \epsilon_{\text{miRanda}} + \epsilon_{\text{intaRNA}}$. 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 sequências 3'UTR de @ genes humanos. Quanto maior for a pontuação s , maior a probabilidade do miRNA em questão interagir somente com *IRF4*.

A terceira pontuação (denominada valor ϵs) foi 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 *IRF4*.

6. RESULTADOS

6.1 SEQUENCIAMENTO DE NOVA GERAÇÃO

Foram realizadas três corridas de sequenciamento, sendo obtidos em 301 ciclos com *reads* do tipo *paired-end*: 22,92 Gbp com 76.397.488 *reads passing filter* na corrida 1, 23,47 Gbp com 78.245.664 *reads passing filter* na corrida 2 e 12.37 Gbp com 41.237.616 *reads* na corrida 3. Cada indivíduo sequenciado apresentou uma porcentagem distinta de contribuição de *reads* identificadas no sequenciamento, mas a diversidade do gene *IRF4* pôde ser avaliada em todos eles.

Com relação à qualidade das *reads* sequenciadas, em todas as corridas as *reads* 1 (*forward*) apresentaram uma porcentagem maior de representantes com Q30 em relação às *reads* 3 (*reverse*), o que se deve ao acúmulo de erros observados à medida que avançam os ciclos de sequenciamento. Em todas as corridas, as *reads* 2 (que identificam os índices de cada *cluster*) apresentaram grande parte das *reads* sequenciadas com Q30, em função do número reduzido de ciclos (8) (Figura 11), o que garante a correta associação de cada *read* à amostra de origem.

	% >= Q30		
	reads 1	reads 2	reads 3
Corrida 1	60,68	92,83	49,76
Corrida 2	63,50	94,44	51,24
Corrida 3	66,65	95,83	57,91

Figura 11: Porcentagem de *reads* com qualidade igual ou superior a 30. Dados obtidos do BaseSpace, Illumina.

6.2 AMOSTRA POPULACIONAL

Foram sequenciadas 228 amostras, sendo as análises que não utilizaram informações sobre haplótipos realizadas a partir deste total e considerando todos os pontos de variação identificados no sequenciamento (105 posições). Quando considerando dados haplotípicos, as análises foram realizadas a partir de um conjunto de 209 indivíduos e 74 pontos de variação. Isso porque, para realizar a inferência pelo PHASE foram descartados os *singletons* (31 posições), uma vez que estes dificultam o desempenho do programa. Adicionalmente, foram excluídas 19 amostras que apresentaram haplótipos inferidos com baixa probabilidade (estipulada neste trabalho como sendo menor que 0,6).

Os indivíduos foram classificados quanto à ancestralidade com base nas informações sobre a origem geográfica de seus antepassados (avós e bisavós). A partir das informações relacionadas aos avós foram definidos 5 grupos ancestrais: euro-descendentes, incluiu os indivíduos que possuíam pelo menos 2 dos antepassados com origem europeia; afro-descendentes, incluiu indivíduos que apresentaram pelo menos 2 dos antepassados com origem africana; asiático-descendentes, incluiu indivíduos com pelo menos 2 dos ancestrais de origem asiática; indígena-descendente, incluiu indivíduos com pelo menos 2 dos antepassados de origem indígena; brasileiros, essa classificação foi dada a indivíduos cuja maior parte dos antepassados fosse de origem brasileira. Considerando a classificação descrita acima, 93 indivíduos foram considerados euro-descendentes, 6 africano-descendentes, 6 asiático-descendentes, 5 indígena-descendente, 109 brasileiros e 9 não puderam ser classificados (indicados como *ni*).

Considerando heteroclassificação baseada nos critérios adotados pelo IBGE, 143 indivíduos foram considerados brancos, 5 amarelos, 49 pardos e 31 indivíduos foram considerados pretos. Segundo a autoclassificação, 140 indivíduos se consideraram brancos, 6 amarelos, 51 pardos, 30 pretos e 1 indivíduo se autodenominou indígena.

Com relação às características de pigmentação, 89 dos indivíduos tinham pele clara, 97 pele média e 42 pele escura. Destes, 22 apresentaram sardas. Em relação à cor dos olhos, 12 indivíduos tinham olhos mel, 26 olhos azuis, 41 olhos verdes, 53 olhos castanho-claros e 96 apresentaram olhos castanho-escuros. Quanto à cor dos cabelos, 2 indivíduos possuíam cabelos ruivos, 10 eram loiro-claros, 26 loiro-escuros, 42 castanho-claros, 84 castanho-escuros e 64 possuíam cabelos pretos. A Figura 12 representa a distribuição das características mencionadas nesta amostra populacional.

Figura 12: Distribuição dos indivíduos quanto às características fenotípicas observadas (GraphPad Prism 5.0). A) distribuição da amostra quanto à ancestralidade. B) distribuição de acordo com a classificação do IBGE (heteroclassificação). C) distribuição de acordo com a presença de sardas. D) distribuição segundo cor de pele. E) distribuição conforme cor dos olhos. F) distribuição de acordo com a cor dos cabelos.

6.3 FREQUÊNCIAS ALÉLICAS, HAPLOTÍPICAS E HETEROZIGOSE

Dos 105 *loci* avaliados, 3 apresentaram mais de 5% de dados incompletos e 4 se mostraram não-polimórficos. A heterozigose média observada total foi $0,11631 \pm 0,15437$ e a heterozigose média esperada foi $0,11949 \pm 0,16081$. Com relação ao Equilíbrio de Hardy-Weinberg, 2 *loci* apresentaram desvios: *locus* 92 ($p=0,04978 \pm 0,00020$) e *locus* 101 ($p=0,04875 \pm 0,00020$). Com relação às frequências alélicas e genotípicas, foram apresentados somente os dados referentes a 18 SNPs, os quais apresentaram forte associação ($OR > 10$) com pelo menos uma das características de pigmentação avaliadas (Tabelas 9,10, 11 e 12). Das análises utilizando dados haplotípicos, foram avaliados 74 *loci*, sendo 2 destes não-polimórficos. Foram identificados 101 haplótipos e a diversidade haplotípica foi de $0,9605 \pm 0,0039$. A distribuição de haplótipos não apresentou desvio em relação ao Equilíbrio de Hardy-Weinberg ($p = 0,8392 \pm 0,0108$).

Tabela 9: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de pele e presença de sardas

SNP		Clara (2n=178)	Média (2n=194)	Escura (2n=84)	com sardas (2n=44)	sem sardas (2n=412)	Total (2n=456)
rs2797305	C/C	0,7045	0,4948	0,2857	0,5909	0,5317	0,5374
	C/T	0,2614	0,3918	0,5714	0,3636	0,3756	0,3744
	T/T	0,0341	0,1134	0,1429	0,0455	0,0927	0,0881
	C	0,8352	0,6907	0,5714	0,7727	0,7195	0,7247
	T	0,1648	0,3093	0,4286	0,2273	0,2805	0,2753
	<i>p</i> EHW	0,6948	0,4763	0,3576	1,0000	0,3008	0,4025
	<i>Ho</i>	0,2614	0,3958	0,5714	0,3636	0,3756	0,3745
	<i>He</i>	0,2755	0,4319	0,4957	0,3594	0,4040	0,3994
rs6929421	C/C	0,9432	0,9485	0,9286	0,9091	0,9463	0,9427
	C/A	0,0568	0,0515	0,0714	0,0909	0,0537	0,0573
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9716	0,9742	0,9643	0,9545	0,9732	0,9714
	A	0,0284	0,0258	0,0357	0,0455	0,0268	0,0286
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	0,0568	0,0521	0,0714	0,0909	0,0537	0,0573
	<i>He</i>	0,0552	0,0510	0,0697	0,0888	0,0522	0,0556
rs77889581	C/C	1,0000	0,9691	0,9048	1,0000	0,9659	0,9692
	C/T	0,0000	0,0309	0,0952	0,0000	0,0341	0,0308
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	1,0000	0,9845	0,9524	1,0000	0,9829	0,9846
	T	0,0000	0,0155	0,0476	0,0000	0,0171	0,0154
	<i>p</i> EHW	-	1,0000	1,0000	-	1,0000	1,0000
	<i>Ho</i>	-	0,0313	0,0952	-	0,0342	0,0308
	<i>He</i>	-	0,0309	0,0918	-	0,0336	0,0304
rs73717071	G/G	1,0000	0,9691	0,9048	1,0000	0,9660	0,9693
	G/A	0,0000	0,0309	0,0952	0,0000	0,0340	0,0307
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	1,0000	0,9845	0,9524	1,0000	0,9830	0,9846
	A	0,0000	0,0155	0,0476	0,0000	0,0170	0,0154
	<i>p</i> EHW	-	1,0000	1,0000	-	1,0000	1,0000
	<i>Ho</i>	-	0,0312	0,0952	-	0,0340	0,0307
	<i>He</i>	-	0,0309	0,0918	-	0,0335	0,0303
rs12203592	C/C	0,7955	0,9381	1,0000	0,7727	0,9073	0,8943
	C/T	0,1591	0,0619	0,0000	0,2273	0,0732	0,0881
	T/T	0,0455	0,0000	0,0000	0,0000	0,0195	0,0176
	C	0,8750	0,9691	1,0000	0,8864	0,9439	0,9383
	T	0,1250	0,0309	0,0000	0,1136	0,0561	0,0617
	<i>p</i> EHW	0,0243	1,0000	-	1,0000	0,0016	0,0059
	<i>Ho</i>	0,1591	0,0625	-	0,22727	0,0732	0,0881
	<i>He</i>	0,2189	0,0609	-	0,20613	0,1059	0,1158
rs41301853	C/C	0,9091	0,9381	0,9286	0,8636	0,9317	0,9251
	C/T	0,0909	0,0619	0,0714	0,1364	0,0683	0,0749
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9545	0,9691	0,9643	0,9318	0,9659	0,9626
	T	0,0455	0,0309	0,0357	0,0682	0,0341	0,0374
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	0,0909	0,0625	0,0714	0,13636	0,0683	0,0749
	<i>He</i>	0,0868	0,0609	0,0697	0,13002	0,0660	0,0721

Continuação Tabela 9: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de pele e presença de sardas

SNP		Clara (2n=178)	Média (2n=194)	Escura (2n=84)	com sardas (2n=44)	sem sardas (2n=412)	Total (2n=456)
rs75719951	C/C	0,9432	0,9381	0,9048	1,0000	0,9268	0,9339
	C/G	0,0568	0,0619	0,0952	0,0000	0,0732	0,0661
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9716	0,9691	0,9524	1,0000	0,9634	0,9670
	G	0,0284	0,0309	0,0476	0,0000	0,0366	0,0330
	pEHW	1,0000	1,0000	1,0000	-	1,0000	1,0000
	Ho	0,0568	0,0625	0,0952	-	0,0732	0,0661
	He	0,0555	0,0609	0,0918	-	0,0707	0,0640
rs113635483	A/A	1,0000	0,9691	0,9524	0,9545	0,9806	0,9781
	A/G	0,0000	0,0309	0,0476	0,0455	0,0194	0,0219
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	1,0000	0,9845	0,9762	0,9773	0,9903	0,9890
	G	0,0000	0,0155	0,0238	0,0227	0,0097	0,0110
	pEHW	-	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	-	0,0313	0,0476	0,0455	0,0194	0,0219
	He	-	0,0309	0,0471	0,0455	0,0193	0,0217
rs1131442	G/G	0,3766	0,3556	0,4412	0,4762	0,3667	0,3781
	G/A	0,4675	0,4667	0,4706	0,4762	0,4667	0,4677
	A/A	0,1558	0,1778	0,0882	0,0476	0,1667	0,1542
	G	0,6104	0,5889	0,6765	0,7143	0,6000	0,6119
	A	0,3896	0,4111	0,3235	0,286	0,4000	0,3881
	pEHW	1,0000	0,6638	1,0000	0,6327	0,7570	0,8830
	Ho	0,4675	0,4607	0,4706	0,4762	0,4667	0,4677
	He	0,4731	0,4847	0,4716	0,4120	0,4817	0,4764
rs4959248	T/T	0,0000	0,0000	0,0238	0,0000	0,0049	0,0044
	T/C	0,0112	0,1134	0,1190	0,1364	0,0680	0,0746
	C/C	0,9888	0,8866	0,8571	0,8636	0,9272	0,9211
	T	0,0056	0,0567	0,0833	0,0682	0,0388	0,0417
	C	0,9944	0,9433	0,9167	0,9318	0,9612	0,9583
	pEHW	1,0000	1,0000	0,2368	1,0000	0,2605	0,3231
	Ho	0,0112	0,1146	0,1191	0,1364	0,0680	0,0746
	He	0,0112	0,1086	0,1546	0,1300	0,0748	0,0800
rs7768807	T/T	0,4944	0,4536	0,4286	0,6818	0,4417	0,4649
	T/C	0,4270	0,4639	0,5000	0,2727	0,4757	0,4561
	C/C	0,0787	0,0825	0,0714	0,0455	0,0825	0,0789
	T	0,7079	0,6856	0,6786	0,8182	0,6796	0,6930
	C	0,2921	0,3144	0,3214	0,1818	0,3204	0,3070
	pEHW	1,0000	0,6372	0,4894	0,5374	0,2041	0,3485
	Ho	0,4269	0,4583	0,5000	0,2727	0,4757	0,4561
	He	0,4159	0,4319	0,4415	0,3044	0,4365	0,4265
rs41301859	T/T	0,9663	0,9485	0,9524	1,0000	0,9515	0,9561
	T/C	0,0337	0,0515	0,0476	0,0000	0,0485	0,0439
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	0,9831	0,9742	0,9762	1,0000	0,9757	0,9781
	C	0,0169	0,0258	0,0238	0,0000	0,0243	0,0219
	pEHW	1,0000	1,0000	1,0000	-	1,0000	1,0000
	Ho	0,0337	0,0521	0,0476	-	0,0485	0,0439
	He	0,0333	0,0510	0,0471	-	0,0475	0,0430

Continuação Tabela 9: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de pele e presença de sardas

SNP		Clara (2n=178)	Média (2n=194)	Escura (2n=84)	com sardas (2n=44)	sem sardas (2n=412)	Total (2n=456)
rs6596980	G/G	0,0000	0,0000	0,0238	0,0000	0,0049	0,0044
	G/A	0,0337	0,1546	0,1190	0,0909	0,1019	0,1009
	A/A	0,9663	0,8454	0,8571	0,9091	0,8932	0,8947
	G	0,0169	0,0773	0,0833	0,0455	0,0558	0,0548
	A	0,9831	0,9227	0,9167	0,9545	0,9442	0,9452
	pEHW	1,0000	1,0000	0,2378	1,0000	0,4785	0,5016
	Ho	0,0337	0,1562	0,1191	0,0909	0,1019	0,1009
	He	0,0333	0,1448	0,1546	0,0888	0,1057	0,1039
rs145752540	G/G	0,9775	0,9588	1,0000	1,0000	0,9709	0,9737
	G/A	0,0225	0,0412	0,0000	0,0000	0,0291	0,0263
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	0,9888	0,9794	1,0000	1,0000	0,9854	0,9868
	A	0,0112	0,0206	0,0000	0,0000	0,0146	0,0132
	pEHW	1,0000	1,0000	-	-	1,0000	1,0000
	Ho	0,0225	0,0417	-	-	0,0291	0,0263
	He	0,0223	0,0410	-	-	0,0288	0,0260
rs78252705	T/T	1,0000	0,9794	0,8571	1,0000	0,9612	0,9649
	T/C	0,0000	0,0206	0,1429	0,0000	0,0388	0,0351
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	1,0000	0,9897	0,9286	1,0000	0,9806	0,9825
	C	0,0000	0,0103	0,0714	0,0000	0,0194	0,0175
	pEHW	-	1,0000	1,0000	-	1,0000	1,0000
	Ho	-	0,0208	0,1429	-	0,0388	0,0351
	He	-	0,0207	0,1343	-	0,0382	0,0345
rs9392502	C/C	0,5955	0,7010	0,7381	0,7727	0,6553	0,6667
	C/G	0,3371	0,2577	0,2619	0,1818	0,3010	0,2895
	G/G	0,0674	0,0412	0,0000	0,0455	0,0437	0,0439
	C	0,7640	0,8299	0,8690	0,8636	0,8058	0,8114
	G	0,2360	0,1701	0,1310	0,1364	0,1942	0,1886
	pEHW	0,5567	0,2861	1,0000	0,3234	0,6537	0,3897
	Ho	0,3371	0,2500	0,2619	0,1818	0,3010	0,2895
	He	0,3626	0,2792	0,2304	0,2410	0,3137	0,3067
rs45580432	A/A	0,9551	0,9485	0,9524	0,9545	0,9515	0,9518
	A/T	0,0449	0,0515	0,0476	0,0455	0,0485	0,0482
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	0,9775	0,9742	0,9762	0,9773	0,9757	0,9759
	T	0,0225	0,0258	0,0238	0,0227	0,0243	0,0241
	pEHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	0,0449	0,0521	0,0476	0,0455	0,0485	0,0483
	He	0,0442	0,0510	0,0471	0,0455	0,0475	0,0472
rs140294761	ATG/ATG	0,8315	0,8763	1,0000	0,9091	0,8786	0,8816
	ATG/A	0,1573	0,1237	0,0000	0,0909	0,1165	0,1140
	AA	0,0112	0,0000	0,0000	0,0000	0,0049	0,0044
	ATG	0,9101	0,9381	1,0000	0,9545	0,9369	0,9386
	A	0,0899	0,0619	0,0000	0,0455	0,0631	0,0614
	pEHW	0,5232	1,0000	-	1,0000	0,5701	0,5856
	Ho	0,1573	0,1250	-	0,0909	0,1165	0,1140
	He	0,1645	0,1178	-	0,0888	0,1185	0,1155

Observação: pEHW: valor de probabilidade de aderência ao Equilíbrio de Hardy-Weinberg; Ho: Heterozigose observada; He: Heterozigose esperada, Desvios em relação ao EHW estão destacados em vermelho, SNPs que foram selecionados por apresentar forte associação com algum fenótipo de cor de pele ou presença de sardas estão destacados em negrito.

Tabela 10: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de olhos

SNP		azul (2n=52)	verde (2n=82)	mel (2n=24)	castanho-claro (2n=106)	castanho-escuro (2n=192)	Total (2n=456)
rs2797305	C/C	0,7200	0,6829	0,5000	0,5849	0,4063	0,5374
	C/T	0,2800	0,2195	0,4167	0,3585	0,4688	0,3744
	T/T	0,0000	0,0976	0,0833	0,0566	0,1250	0,0881
	C	0,8600	0,7927	0,7083	0,7642	0,6406	0,7247
	T	0,1400	0,2073	0,2917	0,2358	0,3594	0,2753
	pEHW	1,0000	0,0454	1,0000	1,0000	1,0000	0,4025
	Ho	0,2800	0,2195	0,4167	0,3585	0,4688	0,3745
	He	0,2416	0,3327	0,4312	0,3639	0,4629	0,3994
rs6929421	C/C	1,0000	0,9268	0,7500	0,9245	0,9688	0,9427
	C/A	0,0000	0,0732	0,2500	0,0755	0,0313	0,0573
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	1,0000	0,9634	0,8750	0,9623	0,9844	0,9714
	A	0,0000	0,0366	0,1250	0,0377	0,0156	0,0286
	pEHW	-	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	-	0,0732	0,2500	0,0755	0,0313	0,0573
	He	-	0,0714	0,2283	0,0733	0,0309	0,0556
rs7788951	C/C	1,0000	0,9756	0,9167	0,9811	0,9583	0,9692
	C/T	0,0000	0,0244	0,0833	0,0189	0,0417	0,0308
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	1,0000	0,9878	0,9583	0,9906	0,9792	0,9846
	T	0,0000	0,0122	0,0417	0,0094	0,0208	0,0154
	pEHW	-	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	-	0,0244	0,0833	0,0189	0,0417	0,0308
	He	-	0,0244	0,0833	0,0189	0,0410	0,0304
rs7371701	G/G	1,0000	0,9756	0,9167	0,9811	0,9583	0,9693
	G/A	0,0000	0,0244	0,0833	0,0189	0,0417	0,0307
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	1,0000	0,9878	0,9583	0,9906	0,9792	0,9846
	A	0,0000	0,0122	0,0417	0,0094	0,0208	0,0154
	pEHW	-	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	-	0,0244	0,0833	0,0189	0,0417	0,0307
	He	-	0,0244	0,0833	0,0189	0,0410	0,0303
rs12203592	C/C	0,6800	0,8780	0,9167	0,8868	0,9583	0,8943
	C/T	0,2400	0,0976	0,0000	0,1132	0,0417	0,0881
	T/T	0,0800	0,0244	0,0833	0,0000	0,0000	0,0176
	C	0,8000	0,9268	0,9167	0,9434	0,9792	0,9383
	T	0,2000	0,0732	0,0833	0,0566	0,0208	0,0617
	pEHW	0,2163	0,1781	0,0434	1,0000	1,0000	0,0059
	Ho	0,2400	0,0976	0,0000	0,1132	0,0417	0,0881
	He	0,3216	0,1373	0,1594	0,1078	0,0410	0,1158
rs4130183	C/C	0,9200	0,9268	0,7500	0,8868	0,9688	0,9251
	C/T	0,0800	0,0732	0,2500	0,1132	0,0313	0,0749
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9600	0,9634	0,8750	0,9434	0,9844	0,9626
	T	0,0400	0,0366	0,1250	0,0566	0,0156	0,0374
	pEHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	0,0800	0,0732	0,2500	0,1132	0,0313	0,0749
	He	0,0769	0,0714	0,2283	0,1078	0,0309	0,0721

Continuação Tabela 10: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de olhos

SNP		azul (2n=52)	verde (2n=82)	mel (2n=24)	castanho-claro (2n=106)	castanho-escuro (2n=192)	Total (2n=456)
rs75719951	C/C	0,9600	0,9268	0,9167	0,9434	0,9271	0,9339
	C/G	0,0400	0,0732	0,0833	0,0566	0,0729	0,0661
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9800	0,9634	0,9583	0,9717	0,9635	0,9670
	G	0,0200	0,0366	0,0417	0,0283	0,0365	0,0330
	pEHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	0,0400	0,0732	0,0833	0,0566	0,0729	0,0661
	He	0,0400	0,0714	0,0833	0,0555	0,0706	0,0640
rs113635483	A/A	1,0000	1,0000	1,0000	0,9434	0,9792	0,9781
	A/G	0,0000	0,0000	0,0000	0,0566	0,0208	0,0219
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	1,0000	1,0000	1,0000	0,9717	0,9896	0,9890
	G	0,0000	0,0000	0,0000	0,0283	0,0104	0,0110
	pEHW	-	-	-	1,0000	1,0000	1,0000
	Ho	-	-	-	0,0566	0,0208	0,0219
	He	-	-	-	0,0555	0,0207	0,0217
rs1131442	G/G	0,3750	0,3333	0,5833	0,3958	0,3580	0,3781
	G/A	0,4583	0,5278	0,3333	0,5208	0,4321	0,4677
	A/A	0,1667	0,1389	0,0833	0,0833	0,2099	0,1542
	G	0,6042	0,5972	0,7500	0,6563	0,5741	0,6119
	A	0,3958	0,4028	0,2500	0,3438	0,4259	0,3881
	pEHW	1,0000	0,7338	1,0000	0,3563	0,3642	0,8830
	Ho	0,4583	0,5278	0,3333	0,5208	0,4321	0,4677
	He	0,4847	0,4819	0,3913	0,4533	0,4944	0,4764
rs4959248	T/T	0,0000	0,0000	0,0000	0,0000	0,0104	0,0044
	T/C	0,0000	0,0244	0,0833	0,0755	0,1146	0,0746
	C/C	1,0000	0,9756	0,9167	0,9245	0,8750	0,9211
	T	0,0000	0,0122	0,0417	0,0377	0,0677	0,0417
	C	1,0000	0,9878	0,9583	0,9623	0,9323	0,9583
	pEHW	-	1,0000	1,0000	1,0000	0,3514	0,3231
	Ho	-	0,0244	0,0833	0,0755	0,1146	0,0746
	He	-	0,0244	0,0833	0,0733	0,1269	0,0800
rs7768807	T/T	0,4615	0,4878	0,5833	0,5283	0,4063	0,4649
	T/C	0,4231	0,3902	0,4167	0,4151	0,5208	0,4561
	C/C	0,1154	0,1220	0,0000	0,0566	0,0729	0,0789
	T	0,6731	0,6829	0,7917	0,7358	0,6667	0,6930
	C	0,3269	0,3171	0,2083	0,2642	0,3333	0,3070
	pEHW	1,0000	0,4900	1,0000	1,0000	0,1123	0,3485
	Ho	0,4231	0,3902	0,4167	0,4151	0,5208	0,4561
	He	0,4487	0,4384	0,3442	0,3925	0,4468	0,4265
rs41301859	T/T	1,0000	0,9512	0,8333	0,9057	0,9896	0,9561
	T/C	0,0000	0,0488	0,1667	0,0943	0,0104	0,0439
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	1,0000	0,9756	0,9167	0,9528	0,9948	0,9781
	C	0,0000	0,0244	0,0833	0,0472	0,0052	0,0219
	pEHW	-	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	-	0,0488	0,1667	0,0943	0,0104	0,0439
	He	-	0,0482	0,1594	0,0908	0,0104	0,0430

Continuação Tabela 10: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de olhos

SNP		azul (2n=52)	verde (2n=82)	mel (2n=24)	castanho-claro (2n=106)	castanho-escuro (2n=192)	Total (2n=456)
rs6596980	G/G	0,0000	0,0000	0,0000	0,0000	0,0104	0,0044
	G/A	0,0000	0,0488	0,2500	0,1698	0,0938	0,1009
	A/A	1,0000	0,9512	0,7500	0,8302	0,8958	0,8947
	G	0,0000	0,0244	0,1250	0,0849	0,0573	0,0548
	A	1,0000	0,9756	0,8750	0,9151	0,9427	0,9452
	pEHW	-	1,0000	1,0000	1,0000	0,2610	0,5016
	Ho	-	0,0488	0,2500	0,1698	0,0938	0,1009
	He	-	0,0482	0,2283	0,1569	0,1086	0,1039
rs145752540	G/G	1,0000	0,9512	1,0000	0,9811	0,9688	0,9737
	G/A	0,0000	0,0488	0,0000	0,0189	0,0313	0,0263
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	1,0000	0,9756	1,0000	0,9906	0,9844	0,9868
	A	0,0000	0,0244	0,0000	0,0094	0,0156	0,0132
	pEHW	-	1,0000	-	1,0000	1,0000	1,0000
	Ho	-	0,0488	-	0,0188	0,0313	0,0263
	He	-	0,0482	-	0,0188	0,0309	0,0260
rs78252705	T/T	1,0000	1,0000	1,0000	1,0000	0,9167	0,9649
	T/C	0,0000	0,0000	0,0000	0,0000	0,0833	0,0351
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	1,0000	1,0000	1,0000	1,0000	0,9583	0,9825
	C	0,0000	0,0000	0,0000	0,0000	0,0417	0,0175
	pEHW	-	-	-	-	1,0000	1,0000
	Ho	-	-	-	-	0,0833	0,0351
	He	-	-	-	-	0,0803	0,0345
rs9392502	C/C	0,5769	0,6585	0,7500	0,6792	0,6771	0,6667
	C/G	0,3462	0,2439	0,2500	0,2642	0,3125	0,2895
	G/G	0,0769	0,0976	0,0000	0,0566	0,0104	0,0439
	C	0,7500	0,7805	0,8750	0,8113	0,8333	0,8114
	G	0,2500	0,2195	0,1250	0,1887	0,1667	0,1886
	pEHW	0,6261	0,0724	1,0000	0,3621	0,4564	0,3897
	Ho	0,3462	0,2439	0,2500	0,2642	0,3125	0,2895
	He	0,3824	0,3469	0,2283	0,3091	0,2792	0,3067
rs45580432	A/A	1,0000	0,9512	0,7500	0,9057	0,9896	0,9518
	A/T	0,0000	0,0488	0,2500	0,0943	0,0104	0,0482
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	1,0000	0,9756	0,8750	0,9528	0,9948	0,9759
	T	0,0000	0,0244	0,1250	0,0472	0,0052	0,0241
	pEHW	-	1,0000	1,0000	1,0000	1,0000	1,0000
	Ho	-	0,0488	0,2500	0,0943	0,0104	0,0483
	He	-	0,0482	0,2283	0,0908	0,0104	0,0472
rs14029471	ATG/ATG	0,8077	0,8293	0,9167	0,9245	0,8958	0,8816
	ATG/A	0,1923	0,1707	0,0000	0,0755	0,1042	0,1140
	A/A	0,0000	0,0000	0,0833	0,0000	0,0000	0,0044
	ATG	0,9038	0,9146	0,9167	0,9623	0,9479	0,9386
	A	0,0962	0,0854	0,0833	0,0377	0,0521	0,0614
	pEHW	1,0000	1,0000	0,0431	1,0000	1,0000	0,5856
	Ho	0,1923	0,1707	0,0000	0,0755	0,1042	0,1140
	He	0,1772	0,1581	0,1594	0,0733	0,0993	0,1155

Observação: pEHW: valor de probabilidade de aderência ao Equilíbrio de Hardy-Weinberg; Ho: Heterozigose observada; He: Heterozigose esperada. Desvios em relação ao EHW estão destacados em vermelho. SNPs que foram selecionados por apresentar forte associação com algum fenótipo de cor dos olhos estão destacados em negrito.

Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo

SNP		ruivo (2n=4)	loiro-claro (2n=10)	loiro-escuro (2n=52)	castanho-claro (2n=84)	castanho-escuro (2n=168)	preto (2n=128)	Total (2n=456)
rs2797305	C/C	0,5000	0,9000	0,6000	0,5714	0,6071	0,3438	0,5374
	C/T	0,5000	0,1000	0,3600	0,4048	0,2857	0,5156	0,3744
	T/T	0,0000	0,0000	0,0400	0,0238	0,1071	0,1406	0,0881
	C	0,7500	0,9500	0,7800	0,7738	0,7500	0,6016	0,7247
	T	0,2500	0,0500	0,2200	0,2262	0,2500	0,3984	0,2753
	<i>p</i> EHW	1,0000	1,0000	1,0000	0,6561	0,0383	0,6129	0,4025
	<i>Ho</i>	0,5000	0,1000	0,3600	0,4048	0,2857	0,5156	0,3745
	<i>He</i>	0,5000	0,1000	0,3451	0,3543	0,3773	0,4831	0,3994
rs6929421	C/C	1,0000	1,0000	0,9200	0,9048	0,9524	0,9531	0,9427
	C/A	0,0000	0,0000	0,0800	0,0952	0,0476	0,0469	0,0573
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	1,0000	1,0000	0,9600	0,9524	0,9762	0,9766	0,9714
	A	0,0000	0,0000	0,0400	0,0476	0,0238	0,0234	0,0286
	<i>p</i> EHW	-	-	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	0,0800	0,0952	0,0476	0,0469	0,0573
	<i>He</i>	-	-	0,0769	0,0918	0,0468	0,0461	0,0556
rs77889581	C/C	1,0000	1,0000	1,0000	0,9762	0,9524	0,9688	0,9692
	C/T	0,0000	0,0000	0,0000	0,0238	0,0476	0,0313	0,0308
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	1,0000	1,0000	1,0000	0,9881	0,9762	0,9844	0,9846
	T	0,0000	0,0000	0,0000	0,0119	0,0238	0,0156	0,0154
	<i>p</i> EHW	-	-	-	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	-	0,0238	0,0476	0,0313	0,0308
	<i>He</i>	-	-	-	0,0238	0,0468	0,0310	0,0304

Continuação Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo

SNP		ruivo (2n=4)	loiro-claro (2n=10)	loiro-escuro (2n=52)	castanho-claro (2n=84)	castanho-escuro (2n=168)	preto (2n=128)	Total (2n=456)
rs73717071	G/G	1,0000	1,0000	1,0000	0,9762	0,9524	0,9688	0,9693
	G/A	0,0000	0,0000	0,0000	0,0238	0,0476	0,0313	0,0307
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	1,0000	1,0000	1,0000	0,9881	0,9762	0,9844	0,9846
	A	0,0000	0,0000	0,0000	0,0119	0,0238	0,0156	0,0154
	<i>p</i> EHW	-	-	-	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	-	0,0238	0,0476	0,0313	0,0307
	<i>He</i>	-	-	-	0,0238	0,0468	0,0310	0,0303
rs12203592	C/C	1,0000	0,9000	0,8800	0,9286	0,8452	0,9375	0,8943
	C/T	0,0000	0,1000	0,0800	0,0714	0,1310	0,0469	0,0881
	T/T	0,0000	0,0000	0,0400	0,0000	0,0238	0,0156	0,0176
	C	1,0000	0,9500	0,9200	0,9643	0,9107	0,9609	0,9383
	T	0,0000	0,0500	0,0800	0,0357	0,0893	0,0391	0,0617
	<i>p</i> EHW	-	1,0000	0,1205	1,0000	0,1177	0,0788	0,0059
	<i>Ho</i>	-	0,1000	0,0800	0,0714	0,1301	0,0469	0,0881
	<i>He</i>	-	0,1000	0,1475	0,0697	0,1636	0,0757	0,1158
rs41301853	C/C	0,5000	0,8000	0,9600	0,9048	0,9405	0,9375	0,9251
	C/T	0,5000	0,2000	0,0400	0,0952	0,0595	0,0625	0,0749
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,7500	0,9000	0,9800	0,9524	0,9702	0,9688	0,9626
	T	0,2500	0,1000	0,0200	0,0476	0,0298	0,0313	0,0374
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	0,5000	0,2000	0,0400	0,0952	0,0595	0,0625	0,0749
	<i>He</i>	0,5000	0,1895	0,0392	0,0918	0,0581	0,0610	0,0721

Continuação Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo

SNP		ruivo (2n=4)	loiro-claro (2n=10)	loiro-escuro (2n=52)	castanho-claro (2n=84)	castanho-escuro (2n=168)	preto (2n=128)	Total (2n=456)
rs75719951	C/C	1,0000	0,9000	0,9600	0,9048	0,9881	0,8750	0,9339
	C/G	0,0000	0,1000	0,0400	0,0952	0,0119	0,1250	0,0661
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	1,0000	0,9500	0,9800	0,9524	0,9940	0,9375	0,9670
	G	0,0000	0,0500	0,0200	0,0476	0,0060	0,0625	0,0330
	<i>p</i> EHW	-	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	0,1000	0,0400	0,0952	0,0119	0,1250	0,0661
	<i>He</i>	-	0,1000	0,0400	0,0918	0,0119	0,1181	0,0640
rs113635483	A/A	1,0000	1,0000	1,0000	0,9762	0,9762	0,9688	0,9781
	A/G	0,0000	0,0000	0,0000	0,0238	0,0238	0,0313	0,0219
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	1,0000	1,0000	1,0000	0,9881	0,9881	0,9844	0,9890
	G	0,0000	0,0000	0,0000	0,0119	0,0119	0,0156	0,0110
	<i>p</i> EHW	-	-	-	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	-	0,0238	0,0238	0,0313	0,0219
	<i>He</i>	-	-	-	0,0238	0,0237	0,0310	0,0217
rs1131442	G/G	0,5000	0,2000	0,3333	0,3514	0,4595	0,3333	0,3781
	G/A	0,5000	0,4000	0,6190	0,5405	0,4189	0,4386	0,4677
	A/A	0,0000	0,4000	0,0476	0,1081	0,1216	0,2281	0,1542
	G	0,7500	0,4000	0,6429	0,6216	0,6689	0,5526	0,6119
	A	0,2500	0,6000	0,3571	0,3784	0,3311	0,4474	0,3881
	<i>p</i> EHW	1,0000	0,5729	0,1837	0,4941	0,6064	0,4262	0,8830
	<i>Ho</i>	0,5000	0,4000	0,6191	0,5405	0,4189	0,4386	0,4677
	<i>He</i>	0,5000	0,5053	0,4493	0,4795	0,4444	0,5014	0,4764

Continuação Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo

SNP		ruivo (2n=4)	loiro-claro (2n=10)	loiro-escuro (2n=52)	castanho-claro (2n=84)	castanho-escuro (2n=168)	preto (2n=128)	Total (2n=456)
rs4959248	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0156	0,0044
	T/C	0,0000	0,0000	0,0385	0,0476	0,0952	0,0938	0,0746
	C/C	1,0000	1,0000	0,9615	0,9524	0,9048	0,8906	0,9211
	T	0,0000	0,0000	0,0192	0,0238	0,0476	0,0625	0,0417
	C	1,0000	1,0000	0,9808	0,9762	0,9524	0,9375	0,9583
	<i>p</i> EHW	-	-	1,0000	1,0000	1,0000	0,2061	0,3231
	<i>Ho</i>	-	-	0,0385	0,0476	0,0952	0,0938	0,0746
	<i>He</i>	-	-	0,0385	0,0471	0,0913	0,1181	0,0800
rs7768807	T/T	0,5000	0,3000	0,5769	0,5000	0,5000	0,3750	0,4649
	T/C	0,5000	0,4000	0,3846	0,4286	0,4167	0,5625	0,4561
	C/C	0,0000	0,3000	0,0385	0,0714	0,0833	0,0625	0,0789
	T	0,7500	0,5000	0,7692	0,7143	0,7083	0,6563	0,6930
	C	0,2500	0,5000	0,2308	0,2857	0,2917	0,3438	0,3070
	<i>p</i> EHW	1,0000	0,5643	1,0000	1,0000	1,0000	0,0934	0,3485
	<i>Ho</i>	0,5000	0,4000	0,3846	0,4286	0,4167	0,5625	0,4561
	<i>He</i>	0,5000	0,5263	0,3611	0,4131	0,4157	0,4547	0,4265
rs41301859	T/T	1,0000	1,0000	0,9615	0,9286	0,9524	0,9688	0,9561
	T/C	0,0000	0,0000	0,0385	0,0714	0,0476	0,0313	0,0439
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	1,0000	1,0000	0,9808	0,9643	0,9762	0,9844	0,9781
	C	0,0000	0,0000	0,0192	0,0357	0,0238	0,0156	0,0219
	<i>p</i> EHW	-	-	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	0,0385	0,0714	0,0476	0,0313	0,0439
	<i>He</i>	-	-	0,0385	0,0697	0,0468	0,0310	0,0430

Continuação Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo

SNP		ruivo (2n=4)	loiro-claro (2n=10)	loiro-escuro (2n=52)	castanho-claro (2n=84)	castanho-escuro (2n=168)	preto (2n=128)	Total (2n=456)
rs6596980	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0156	0,0044
	G/A	0,0000	0,0000	0,0385	0,1190	0,1310	0,0938	0,1009
	A/A	1,0000	1,0000	0,9615	0,8810	0,8690	0,8906	0,8947
	G	0,0000	0,0000	0,0192	0,0595	0,0655	0,0625	0,0548
	A	1,0000	1,0000	0,9808	0,9405	0,9345	0,9375	0,9452
	<i>p</i> EHW	-	-	1,0000	1,0000	1,0000	0,2078	0,5016
	<i>Ho</i>	-	-	0,0385	0,1191	0,1301	0,0938	0,1009
	<i>He</i>	-	-	0,0385	0,1133	0,1231	0,1181	0,1039
rs145752540	G/G	1,0000	1,0000	0,9615	0,9524	1,0000	0,9531	0,9737
	G/A	0,0000	0,0000	0,0385	0,0476	0,0000	0,0469	0,0263
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	1,0000	1,0000	0,9808	0,9762	1,0000	0,9766	0,9868
	A	0,0000	0,0000	0,0192	0,0238	0,0000	0,0234	0,0132
	<i>p</i> EHW	-	-	1,0000	1,0000	-	1,0000	1,0000
	<i>Ho</i>	-	-	0,0385	0,0476	-	0,0469	0,0263
	<i>He</i>	-	-	0,0385	0,0471	-	0,0461	0,0260
rs78252705	T/T	1,0000	1,0000	1,0000	1,0000	0,9762	0,9063	0,9649
	T/C	0,0000	0,0000	0,0000	0,0000	0,0238	0,0938	0,0351
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	1,0000	1,0000	1,0000	1,0000	0,9881	0,9531	0,9825
	C	0,0000	0,0000	0,0000	0,0000	0,0119	0,0469	0,0175
	<i>p</i> EHW	-	-	-	-	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	-	-	0,0238	0,0938	0,0351
	<i>He</i>	-	-	-	-	0,0237	0,0901	0,0345

Continuação Tabela 11: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de cor de cabelo

SNP		ruivo (2n=4)	loiro-claro (2n=10)	loiro-escuro (2n=52)	castanho-claro (2n=84)	castanho-escuro (2n=168)	preto (2n=128)	Total (2n=456)
rs9392502	C/C	0,5000	0,4000	0,7308	0,6905	0,6429	0,7031	0,6667
	C/G	0,5000	0,3000	0,2692	0,2381	0,3214	0,2813	0,2895
	G/G	0,0000	0,3000	0,0000	0,0714	0,0357	0,0156	0,0439
	C	0,7500	0,5500	0,8654	0,8095	0,8036	0,8438	0,8114
	G	0,2500	0,4500	0,1346	0,1905	0,1964	0,1563	0,1886
	<i>p</i> EHW	1,0000	0,2453	1,0000	0,1406	1,0000	1,0000	0,3897
	<i>Ho</i>	0,5000	0,3000	0,2692	0,2381	0,3214	0,2813	0,2895
	<i>He</i>	0,5000	0,5211	0,2376	0,3121	0,3176	0,2658	0,3067
rs45580432	A/A	1,0000	1,0000	0,9615	0,9048	0,9524	0,9688	0,9518
	A/T	0,0000	0,0000	0,0385	0,0952	0,0476	0,0313	0,0482
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	1,0000	1,0000	0,9808	0,9524	0,9762	0,9844	0,9759
	T	0,0000	0,0000	0,0192	0,0476	0,0238	0,0156	0,0241
	<i>p</i> EHW	-	-	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	-	-	0,0385	0,0952	0,0476	0,0313	0,0483
	<i>He</i>	-	-	0,0385	0,0918	0,0468	0,0310	0,0472
rs140294761	ATG/ATG	0,5000	0,7000	0,8462	0,8571	0,9048	0,9219	0,8816
	ATG/A	0,5000	0,3000	0,1538	0,1190	0,0952	0,0781	0,1140
	A/A	0,0000	0,0000	0,0000	0,0238	0,0000	0,0000	0,0044
	ATG	0,7500	0,8500	0,9231	0,9167	0,9524	0,9609	0,9386
	A	0,2500	0,1500	0,0769	0,0833	0,0476	0,0391	0,0614
	<i>p</i> EHW	1,0000	1,0000	1,0000	0,2361	1,0000	1,0000	0,5856
	<i>Ho</i>	0,5000	0,3000	0,1539	0,1191	0,0952	0,0781	0,1140
	<i>He</i>	0,5000	0,2684	0,1448	0,1546	0,0913	0,0757	0,1155

Observação: *p*EHW: valor de probabilidade de aderência ao Equilíbrio de Hardy-Weinberg; *Ho*: Heterozigose observada; *He*: Heterozigose esperada. Desvios em realação ao EHW estão destacados em vermelho. SNPs que foram selecionados por apresentar forte associação com algum fenótipo de cor dos cabelos estão destacados em negrito.

Tabela 12: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de presença de cabelos brancos

SNP		0% (2n=232)	0-25% (2n=82)	25-50% (2n=40)	50-75% (2n=32)	75-100% (2n=22)	Total (2n=456)
rs2797305	C/C	0,5172	0,5750	0,4500	0,8750	0,5455	0,5374
	C/T	0,3966	0,3250	0,4000	0,1250	0,2727	0,3744
	T/T	0,0862	0,1000	0,1500	0,0000	0,1818	0,0881
	C	0,7155	0,7375	0,6500	0,9375	0,6818	0,7247
	T	0,2845	0,2625	0,3500	0,0625	0,3182	0,2753
	<i>p</i> EHW	0,8198	0,4085	0,6337	1,0000	0,2304	0,4025
	<i>H_o</i>	0,3966	0,3250	0,4000	0,1250	0,2727	0,3745
	<i>He</i>	0,4089	0,3889	0,4667	0,1210	0,4546	0,3994
rs6929421	C/C	0,9483	0,9500	0,9500	0,9375	0,8182	0,9427
	C/A	0,0517	0,0500	0,0500	0,0625	0,1818	0,0573
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9741	0,9750	0,9750	0,9688	0,9091	0,9714
	A	0,0259	0,0250	0,0250	0,0313	0,0909	0,0286
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>H_o</i>	0,0517	0,0500	0,0500	0,0625	0,1818	0,0573
	<i>He</i>	0,0506	0,0488	0,0500	0,0625	0,1732	0,0556
rs77889581	C/C	0,9483	0,9750	1,0000	1,0000	1,0000	0,9692
	C/T	0,0517	0,0250	0,0000	0,0000	0,0000	0,0308
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9741	0,9875	1,0000	1,0000	1,0000	0,9846
	T	0,0259	0,0125	0,0000	0,0000	0,0000	0,0154
	<i>p</i> EHW	1,0000	1,0000	-	-	-	1,0000
	<i>H_o</i>	0,0517	0,0250	-	-	-	0,0308
	<i>He</i>	0,0506	0,0247	-	-	-	0,0304
rs73717071	G/G	0,9483	0,9756	1,0000	1,0000	1,0000	0,9693
	G/A	0,0517	0,0244	0,0000	0,0000	0,0000	0,0307
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	0,9741	0,9878	1,0000	1,0000	1,0000	0,9846
	A	0,0259	0,0122	0,0000	0,0000	0,0000	0,0154
	<i>p</i> EHW	1,0000	1,0000	-	-	-	1,0000
	<i>H_o</i>	0,0517	0,0244	-	-	-	0,0307
	<i>He</i>	0,0506	0,0244	-	-	-	0,0303
rs12203592	C/C	0,9138	0,8500	0,9500	0,8750	0,7273	0,8943
	C/T	0,0776	0,1500	0,0500	0,0625	0,0909	0,0881
	T/T	0,0086	0,0000	0,0000	0,0625	0,1818	0,0176
	C	0,9526	0,9250	0,9750	0,9063	0,7727	0,9383
	T	0,0474	0,0750	0,0250	0,0938	0,2273	0,0617
	<i>p</i> EHW	0,2194	1,0000	1,0000	0,0967	0,0376	0,0059
	<i>H_o</i>	0,0776	0,1500	0,0500	0,0625	0,0909	0,0881
	<i>He</i>	0,0907	0,1389	0,0500	0,1754	0,3680	0,1158

Continuação Tabela 12: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de presença de cabelos brancos

SNP		0% (2n=232)	0-25% (2n=82)	25-50% (2n=40)	50-75% (2n=32)	75-100% (2n=22)	Total (2n=456)
rs41301853	C/C	0,9397	0,9250	0,9000	0,8750	0,8182	0,9251
	C/T	0,0603	0,0750	0,1000	0,1250	0,1818	0,0749
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9698	0,9625	0,9500	0,9375	0,9091	0,9626
	T	0,0302	0,0375	0,0500	0,0625	0,0909	0,0374
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	0,0603	0,0750	0,1000	0,1250	0,1818	0,0749
	<i>He</i>	0,0588	0,0722	0,0974	0,1210	0,1732	0,0721
rs75719951	C/C	0,9655	0,9000	0,8000	1,0000	0,9091	0,9339
	C/G	0,0345	0,1000	0,2000	0,0000	0,0909	0,0661
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	C	0,9828	0,9500	0,9000	1,0000	0,9545	0,9670
	G	0,0172	0,0500	0,1000	0,0000	0,0455	0,0330
	<i>p</i> EHW	1,0000	1,0000	1,0000	-	1,0000	1,0000
	<i>Ho</i>	0,0345	0,1000	0,2000	-	0,0909	0,0661
	<i>He</i>	0,0340	0,0962	0,1846	-	0,0909	0,0640
rs113635483	A/A	0,9914	1,0000	0,9500	0,9375	0,8182	0,9781
	A/G	0,0086	0,0000	0,0500	0,0625	0,1818	0,0219
	G/G	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	0,9957	1,0000	0,9750	0,9688	0,9091	0,9890
	G	0,0043	0,0000	0,0250	0,0313	0,0909	0,0110
	<i>p</i> EHW	1,0000	-	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	0,0086	-	0,0500	0,0625	0,1818	0,0219
	<i>He</i>	0,0086	-	0,0500	0,0625	0,1732	0,0217
rs1131442	G/G	0,3810	0,3333	0,4211	0,4286	0,5000	0,3781
	G/A	0,4476	0,5152	0,3158	0,5714	0,5000	0,4677
	A/A	0,1714	0,1515	0,2632	0,0000	0,0000	0,1542
	G	0,6048	0,5909	0,5789	0,7143	0,7500	0,6119
	A	0,3952	0,4091	0,4211	0,2857	0,2500	0,3881
	<i>p</i> EHW	0,5425	1,0000	0,1615	0,5050	1,0000	0,8830
	<i>Ho</i>	0,4476	0,5152	0,3158	0,5714	0,5000	0,4677
	<i>He</i>	0,4796	0,4909	0,5047	0,4138	0,4561	0,4764
rs4959248	T/T	0,0000	0,0000	0,0000	0,0000	0,0909	0,0044
	T/C	0,0517	0,0976	0,0500	0,0625	0,0909	0,0746
	C/C	0,9483	0,9024	0,9500	0,9375	0,8182	0,9211
	T	0,0259	0,0488	0,0250	0,0313	0,1364	0,0417
	C	0,9741	0,9512	0,9750	0,9688	0,8636	0,9583
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	0,1429	0,3231
	<i>Ho</i>	0,0517	0,0976	0,0500	0,0625	0,0909	0,0746
	<i>He</i>	0,0506	0,0940	0,0500	0,0625	0,2468	0,0800

Continuação Tabela 12: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de presença de cabelos brancos

SNP		0% (2n=232)	0-25% (2n=82)	25-50% (2n=40)	50-75% (2n=32)	75-100% (2n=22)	Total (2n=456)
rs7768807	T/T	0,4310	0,4878	0,5000	0,5625	0,3636	0,4649
	T/C	0,4914	0,4146	0,4500	0,3750	0,5455	0,4561
	C/C	0,0776	0,0976	0,0500	0,0625	0,0909	0,0789
	T	0,6767	0,6951	0,7250	0,7500	0,6364	0,6930
	C	0,3233	0,3049	0,2750	0,2500	0,3636	0,3070
	<i>p</i> EHW	0,2876	1,0000	1,0000	1,0000	1,0000	0,3485
	<i>Ho</i>	0,4914	0,4146	0,4500	0,3750	0,5455	0,4561
	<i>He</i>	0,43943	0,4291	0,4090	0,3871	0,4849	0,4265
rs41301859	T/T	0,9655	0,9268	0,9500	0,9375	0,9091	0,9561
	T/C	0,0345	0,0732	0,0500	0,0625	0,0909	0,0439
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	0,9828	0,9634	0,9750	0,9688	0,9545	0,9781
	C	0,0172	0,0366	0,0250	0,0313	0,0455	0,0219
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>Ho</i>	0,0345	0,0732	0,0500	0,0625	0,0909	0,0439
	<i>He</i>	0,0340	0,0714	0,0500	0,0625	0,0909	0,0430
rs6596980	G/G	0,0000	0,0000	0,0000	0,0000	0,0909	0,0044
	G/A	0,0690	0,1220	0,1000	0,1250	0,1818	0,1009
	A/A	0,9310	0,8780	0,9000	0,8750	0,7273	0,8947
	G	0,0345	0,0610	0,0500	0,0625	0,1818	0,0548
	A	0,9655	0,9390	0,9500	0,9375	0,8182	0,9452
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	0,2793	0,5016
	<i>Ho</i>	0,0690	0,1220	0,1000	0,1250	0,1818	0,1009
	<i>He</i>	0,0669	0,1159	0,0974	0,1210	0,3117	0,1039
rs145752540	G/G	0,9828	0,9756	0,8500	1,0000	1,0000	0,9737
	G/A	0,0172	0,0244	0,1500	0,0000	0,0000	0,0263
	A/A	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	G	0,9914	0,9878	0,9250	1,0000	1,0000	0,9868
	A	0,0086	0,0122	0,0750	0,0000	0,0000	0,0132
	<i>p</i> EHW	1,0000	1,0000	1,0000	-	-	1,0000
	<i>Ho</i>	0,0172	0,0244	0,1500	-	-	0,0263
	<i>He</i>	0,0172	0,0244	0,1423	-	-	0,0260
rs78252705	T/T	0,9655	0,9268	1,0000	1,0000	1,0000	0,9649
	T/C	0,0345	0,0732	0,0000	0,0000	0,0000	0,0351
	C/C	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	T	0,9828	0,9634	1,0000	1,0000	1,0000	0,9825
	C	0,0172	0,0366	0,0000	0,0000	0,0000	0,0175
	<i>p</i> EHW	1,0000	1,0000	-	-	-	1,0000
	<i>Ho</i>	0,0345	0,0732	-	-	-	0,0351
	<i>He</i>	0,0340	0,0714	-	-	-	0,0345

Continuação Tabela 12: Frequências alélicas e genotípicas e parâmetros de diversidade intrapopulacional em relação aos grupos de presença de cabelos brancos

SNP		0% (2n=232)	0-25% (2n=82)	25-50% (2n=40)	50-75% (2n=32)	75-100% (2n=22)	Total (2n=456)
rs9392502	C/C	0,6724	0,6829	0,6500	0,6875	0,4545	0,6667
	C/G	0,2845	0,2439	0,3000	0,3125	0,4545	0,2895
	G/G	0,0431	0,0732	0,0500	0,0000	0,0909	0,0439
	C	0,8147	0,8049	0,8000	0,8438	0,6818	0,8114
	G	0,1853	0,1951	0,2000	0,1563	0,3182	0,1886
	<i>p</i> EHW	0,5379	0,1508	1,0000	1,0000	1,0000	0,3897
	<i>H</i> _o	0,2845	0,2439	0,3000	0,3125	0,4545	0,2895
	<i>H</i> _e	0,3033	0,3179	0,3282	0,2722	0,4545	0,3067
rs45580432	A/A	0,9655	0,9268	0,9500	0,9375	0,9091	0,9518
	A/T	0,0345	0,0732	0,0500	0,0625	0,0909	0,0482
	T/T	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	A	0,9828	0,9634	0,9750	0,9688	0,9545	0,9759
	T	0,0172	0,0366	0,0250	0,0313	0,0455	0,0241
	<i>p</i> EHW	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
	<i>H</i> _o	0,0345	0,0732	0,0500	0,0625	0,0909	0,0483
	<i>H</i> _e	0,0340	0,0714	0,0500	0,0625	0,0909	0,0472
rs140294761	ATG/ATG	0,8621	0,9268	0,9000	0,8750	1,0000	0,8816
	ATG/A	0,1293	0,0732	0,1000	0,1250	0,0000	0,1140
	A/A	0,0086	0,0000	0,0000	0,0000	0,0000	0,0044
	ATG	0,9267	0,9634	0,9500	0,9375	1,0000	0,9386
	A	0,0733	0,0366	0,0500	0,0625	0,0000	0,0614
	<i>p</i> EHW	0,4673	1,0000	1,0000	1,0000	-	0,5856
	<i>H</i> _o	0,1293	0,0732	0,1000	0,1250	-	0,1140
	<i>H</i> _e	0,1364	0,0714	0,0974	0,1210	-	0,1155

Observação: *p*EHW: valor de probabilidade de aderência ao Equilíbrio de Hardy-Weinberg; *H*_o: Heterozigose observada; *H*_e: Heterozigose esperada. Desvios em relação ao EHW estão destacados em vermelho. SNPs que foram selecionados por apresentar forte associação com algum fenótipo de cor dos olhos estão destacados em negrito.

6.4 ASSOCIAÇÕES ENCONTRADAS

A fim de relatar apenas as associações mais fortes dentre as associações alélicas e genótípicas encontradas, foram selecionados resultados que apresentassem valores de OR superiores a 10 (com limite inferior do intervalo de confiança de 95% superior a 0,9) e valores de probabilidade obtidos por meio de Teste Exato de Fisher bilateral inferiores a 0,05, exceto para a característica “presença de sardas”, que teve como critério diferenciado valores de OR maiores que 1 (Tabela 13).

Tabela 13: Alelos e genótipos de SNPs/indels do gene *IRF4* associados a características de pigmentação

SNPs	Alelo/ Genótipo	Características	valor de <i>p</i>	Odds ratio (95% IC)
rs2797305	C	cabelo loiro-claro x cabelo preto	0,0019	12,5844 (1,6334-96,9539)
	C/C	cabelo loiro-claro x cabelo preto	0,0013	17,1818 (2,0431-144,4966)
rs6929421	C/C	olho castanho-escuro x olho mel	0,0176	10,3333 (1,8129-58,8981)
rs77889581	C	pele clara x pele escura	0,0104	19,7329 (1,0499-370,8868)
	C/C	pele clara x pele escura	0,0099	20,6883 (1,0870-393,7484)
rs73717071	G	pele clara x pele escura	0,0101	19,9565 (1,0618-375,0739)
	G/G	pele clara x pele escura	0,0096	20,9221 (1,0994-398,1647)
rs12203592	T	pele clara x pele escura	0,0002	24,6117 (1,4744-410,8283)
	T	olho azul x olho castanho-escuro	0,0000	11,7500 (3,5083-39,3527)
	C/C	pele escura x pele clara	0,0007	22,3050 (1,3101-379,7429)
	C/C	olho castanho-escuro x olho azul	0,0003	10,8235 (2,9292-39,9940)
	C/T	pele clara x pele escura	0,0047	16,5436 (0,9625-284,3647)
	C/T	presença de sardas	0,0312	3,7255 (1,2068-11,5006)
	T/T	olho azul x olho castanho-escuro	0,0413	20,5319 (0,9534-442,1435)
	T/T	olho azul x olhos castanhos	0,0199	31,8085 (1,4805-683,4054)
	T/T	acima de 50% de cabelos brancos	0,0214	14,3750 (1,4333-144,1721)
	T/T	acima de 75% de cabelos brancos	0,0196	25,5556 (2,1093-309,6160)
rs41301853	C/C	olho castanho-escuro x olho mel	0,0176	10,3333 (1,8129-58,8981)
rs75719951	C	cabelo castanho-escuro x preto	0,0117	11,1333 (1,3742-90,1964)
	C/C	cabelo castanho-escuro x preto	0,0105	11,8571 (1,4429-97,4399)
rs113635483	A/G	acima de 50% de cabelos brancos	0,0214	14,3750 (1,4333-144,1721)
	A	acima de 75% de cabelos brancos	0,0204	23,1000 (2,0063-265,9673)
	A/G	acima de 75% de cabelos brancos	0,0196	25,5556 (2,1093-309,6160)
rs1131442	A/A	cabelo loiro-claro x loiro-escuro	0,0274	13,3333 (1,2419-143,1512)
rs4959248	C	pele clara x pele escura	0,0017	16,0909 (1,9463-133,0279)
	C/C	pele clara x pele escura	0,0044	14,6667 (1,7046-126,1914)
	C	pele clara x pele média	0,0061	10,6393 (1,3594-83,2690)
	C/C	pele clara x pele média	0,0054	11,2558 (1,4223-89,0733)
rs7768807	T/T	presença de sardas	0,0422	2,7080 (1,0596-6,9210)
rs41301859	C	olho mel x olho castanho-escuro	0,0332	17,3636 (1,5124-199,3490)
	T/T	olho castanho-escuro x olho mel	0,0321	19,0000 (1,5795-228,5514)
rs6596980	A/A	olho azul x olho mel	0,0261	19,5263 (0,9206-414,1530)
rs145752540	G/A	entre 25-50% de cabelos brancos	0,0227	10,0588 (1,5653-64,6375)
rs78252705	T	pele clara x pele escura	0,0010	29,5605 (1,6451-531,1802)
	T/T	pele clara x pele escura	0,0008	31,8767 (1,7503-580,5392)
rs9392502	G/G	cabelo loiro-claro x castanho-escuro	0,0150	11,5714 (1,9576-68,4008)
	G/G	cabelo loiro-claro x loiro-escuro	0,0168	24,7333 (1,1458-533,9087)
	G/G	cabelo loiro-claro x preto	0,0069	27,0000 (2,4639-295,8741)
	G/G	olho verde x olho castanho-escuro	0,0280	10,2703 (1,1110-94,9388)
rs45580432	T	olho mel x olho castanho-escuro	0,0045	27,2857 (2,7149-274,2269)
	A/T	olho mel x olho castanho-escuro	0,0040	31,6667 (2,9776-336,7752)
	A/A	olho castanho-escuro x olho mel	0,0040	31,6667 (2,9776-336,7752)
	A/A	olho azul x olho mel	0,0261	19,5263 (0,9206-414,1530)

Continuação Tabela 13: Alelos e genótipos de SNPs/indels do gene *IRF4* associados a características de pigmentação

SNPs	Alelo/ Genótipo	Características	valor de <i>p</i>	<i>Odds ratio</i> (95% IC)
rs140294761	A	pele clara x pele escura	0,0020	17,1600 (1,0170-289,5494)
	ATG/A	pele clara x pele escura	0,0048	16,3245 (0,9498-280,5646)
	ATG/ATG	pele escura x pele clara	0,0026	17,6846 (1,0319-303,0636)

Observação: Os valores em negrito permanecem significativos ao nível de 5% mesmo se a correção de Bonferroni for levada em consideração. Considerando que cada um dos 105 SNPs possui apenas dois alelos, a informação de cada um dos dois torna-se redundante. Portanto, para alelos, foi considerado $\alpha_{\text{corrigido}} = 4,76 \times 10^{-4}$ (isto é, 0,05/105). Considerando que cada um dos 105 SNPs possui apenas três genótipos, foi considerado $\alpha_{\text{corrigido}} = 1,59 \times 10^{-4}$ (isto é, 0,05/315).

É importante ressaltar que se a extremamente conservadora correção de Bonferroni for levada em consideração, apenas duas associações, envolvendo o SNP rs12203592, permanecem significantes ao nível de 5%.

Com relação às associações encontradas envolvendo haplótipos, divididos de acordo com a região gênica, foram mantidos resultados que apresentassem valores de OR superiores a 1 (com limite inferior do intervalo de confiança de 95% superior a 0,9) e valores de probabilidade obtidos por meio de Teste Exato de Fisher bilateral inferiores a 0,05 (Tabela 14). É importante ressaltar que se a extremamente conservadora correção de Bonferroni for considerada, nenhum valor permanece significativo ao nível de 5%.

Tabela 14: Haplótipos formados por SNPs/indels do gene *IRF4* associados a características de pigmentação

Haplótipo	Características	valor de <i>p</i>	<i>Odds Ratio</i> (95% IC)
PROMOTORA			
PROMO001	pele clara x pele escura	0,0024	2,3989 (1,3609 - 4,2288)
PROMO001	pele clara x pele média	0,0065	1,8305 (1,1884 - 2,8196)
PROMO001	olhos azuis x castanho-escuros	0,0160	2,3183 (1,1841 - 4,5390)
PROMO001	olhos azuis x castanhos	0,0318	2,0606 (1,0818 - 3,9250)
PROMO001	cabelos castanho-escuros x pretos	0,0194	1,8095 (1,1067 - 2,9588)
PROMO001	cabelos castanho-escuros x outros	0,0247	1,5882 (1,0631 - 2,3728)
PROMO001	50-75% x sem cabelos brancos	0,0336	2,3276 (1,0905 - 4,9682)
PROMO001	acima de 50% x sem cabelos brancos	0,0284	2,0081 (1,0870 - 3,7096)
PROMO002	cabelos loiro-claros x pretos	0,0479	2,9333 (1,0656 - 8,0747)
PROMO003	25-50% x sem cabelos brancos	0,0378	3,6022 (1,1329 - 11,4537)
PROMO006	olhos mel x castanho-escuros	0,0247	8,1429 (1,5433 - 42,9648)
CD			
CD001	pele clara x pele escura	0,0034	2,3182 (1,3363 - 4,0215)
CD001	pele média x pele escura	0,0147	1,9614 (1,1437 - 3,3638)
CD001	cabelos loiro-claros x pretos	0,0099	4,5980 (1,4280 - 14,8054)
CD001	cabelos loiro-escuros x pretos	0,0035	2,8902 (1,4199 - 5,8830)
CD001	cabelos castanho-claros x pretos	0,0135	2,1020 (1,1725 - 3,7682)
CD001	cabelos castanho-escuros x pretos	0,0069	2,0029 (1,2310 - 3,2588)
CD002	cabelos castanho-escuros x todos os outros	0,0413	2,2662 (1,0318 - 4,9775)
CD002	pele clara x pele escura	0,0003	23,8551 (1,4256 - 399,1698)
CD002	pele clara x pele média	0,0011	4,2695 (1,6775 - 10,8667)
CD002	olhos azuis x castanho-escuros	0,0001	10,9286 (3,1857 - 37,4902)
CD002	olhos azuis x castanho-claros	0,0165	3,9429 (1,3074 - 11,8912)
CD002	olhos azuis x castanhos	0,0003	6,7371 (2,5612 - 17,7218)
CD002	olhos azuis x verdes	0,0482	3,0857 (1,0178 - 9,3551)
CD002	olhos azuis x todos menos azuis	0,0008	5,0857 (2,1259 - 12,1663)

Continuação Tabela 14: Haplótipos formados por SNPs/indels do gene *IRF4* associados a características de pigmentação

Haplótipo	Características	valor de <i>p</i>	Odds Ratio (95% IC)
CD			
CD002	75-100% x sem cabelos brancos	0,0051	6,6667 (2,0183 - 22,0211)
CD002	acima de 50% x sem cabelos brancos	0,0124	3,6364 (1,3575 - 9,7405)
CD004	olhos mel x castanho-escuros	0,0247	8,1429 (1,5433 - 42,9648)
CD005	cabelos castanho-claros x castanho-escuros	0,0449	8,2703 (0,9083 - 75,2998)
CD005	25-50% x sem cabelos brancos	0,0177	6,4375 (1,5328 - 27,0371)
CD011	cabelos loiro-escuros x castanho-escuros	0,0424	10,2000 (1,0356 - 100,4676)
CD013	acima de 50% x sem cabelos brancos	0,0256	12,7959 (1,3030 - 125,6598)
CDS			
CDS001	pele clara x pele escura	0,0396	8,7027 (0,9561 - 79,2147)
CDS001	pele clara x pele média	0,0210	8,5740 (1,0741 - 68,4401)
3'UTR			
UTR001	pele clara x pele escura	0,0088	2,6966 (1,2782 - 5,6887)
UTR001	pele média x pele escura	0,0445	2,1659 (1,0259 - 4,5728)
UTR001	olhos azuis x castanho-escuros	0,0288	2,2694 (1,1075 - 4,6503)
UTR002	cabelos loiro-claros x pretos	0,0356	3,4333 (1,1205 - 10,5196)
UTR005	50-75% x sem cabelos brancos	0,0071	5,1538 (1,6971 - 15,6515)
UTR007	pele clara x pele escura	0,0010	29,1379 (1,6198 - 524,1460)
UTR009	pele clara x pele média	0,0076	16,2023 (0,9277 - 282,9882)
UTR011	25-50% x sem cabelos brancos	0,0235	9,4545 (1,5220 - 58,7325)

Com relação às associações encontradas envolvendo os haplótipos considerando a região total do gene, foram mantidos resultados que apresentassem valores de OR superiores a 1 (com limite inferior do intervalo de confiança de 95% superior a 0,9) e valores de probabilidade obtidos por meio de Teste Exato de Fisher bilateral inferiores a 0,05 (Tabela 15).

Tabela 15: Haplótipos formados por SNPs de *IRF4* associados a características de pigmentação

Haplótipo	Características de pigmentação	valor de <i>p</i>	Odds Ratio (95% IC)
hp006	cabelos castanho-escuros x pretos	0,0478	2,1938 (1,0127-4,7521)
hp007	cabelos loiro-claros x castanho-escuros	0,0198	4,5513 (1,3876-14,9281)
hp007	cabelos loiro-claros x castanho-claros	0,0300	4,6154 (1,2260-17,3755)
hp007	cabelos loiro-claros x pretos	0,0039	8,6923 (2,2175-34,0722)
hp015	pele clara x pele média	0,0034	3,8089 (1,4815-9,7922)
hp015	pele clara x pele escura	0,0006	21,335 (1,2709-358,1301)
hp015	olhos azuis + verdes x olhos castanhos	0,0097	3,1248 (1,3301-7,3412)
hp015	olhos azuis x não-azuis	0,0099	3,7417 (1,4669-9,5442)
hp015	olhos azuis x olhos castanhos	0,0042	4,9568 (1,7778-13,8203)
hp015	olhos azuis x olhos castanho-escuros	0,0015	8,0405 (2,2380-28,8869)
hp019	pele clara x pele média	0,0155	9,1948 (1,1372-74,3426)
hp041	olhos azuis x olhos castanho-escuros	0,0400	20,5294 (0,9676-435,5784)
hp041	olhos azuis x olhos castanhos	0,0190	32,0588 (1,5129-679,3202)
hp041	olhos azuis x não-azuis	0,0109	44,0588 (2,0806-933,0111)
hp056	cabelos loiro-escuros x castanho-escuros	0,0424	10,2000 (1,0356-100,4676)
hp057	cabelos ruivos x loiro-escuros	0,0400	97,0000 (2,6785-3512,7321)
hp077	olhos azuis x olhos castanho-escuros	0,0400	20,5294 (0,9676-435,5784)
hp077	olhos azuis x castanhos	0,0190	32,0588 (1,5129-679,3202)
hp077	olhos azuis x não-azuis	0,0109	44,0588 (2,0806-933,0111)
hp077	cabelos loiro-claros x castanho-claros	0,0336	23,7879 (1,0906-518,8352)
hp077	cabelos loiro-claros x castanho-escuros	0,0104	46,8182 (2,1550-1017,1650)
hp077	cabelos loiro-claros x pretos	0,0167	35,9091 (1,6508-781,1132)

6.5 DESEQUILÍBRIO DE LIGAÇÃO

O desequilíbrio de ligação entre os *loci* estudados foi avaliado através do programa Haploview 4.2. O arquivo de entrada para o programa foi construído a partir dos dados contendo fase gamética conhecida (209 indivíduos) e foi composto por 74 marcadores bialélicos.

Uma medida do DL entre dois blocos pode ser representada através do valor D' . Quanto mais próximo de zero for esse valor, maior é a taxa de recombinação entre dois blocos, ou seja, menor o DL (Figura 13).

Figura 13: D' e frequências dos haplótipos inferidos em cada bloco pelo Haploview. Os números indicando os SNPs são mostrados acima. Todos os SNPs indicados nesta figura estão realçados em negrito na Figura 14, e são aqueles com DL mais intenso dentro de um bloco. Os *tag*SNPs estão destacados com uma seta verde. As frequências populacionais para cada haplótipo são indicadas ao lado destes e as linhas entre os haplótipos de cada bloco indicam através da espessura uma maior ou menor ligação entre os blocos avaliados. O valor de D' é mostrado na parte inferior.

A associação não-aleatória entre alelos de diferentes *loci* foi quantificada pelo Haploview através de três medidas: o LOD (*log of the likelihood odds ratio*), a medida r^2 e o D' . A Figura 13 representa através das cores o DL entre os SNPs avaliados, sendo o branco representativo de valores de $\text{LOD} < 2$ e $D' < 1$ (DL fraco); azul indica $\text{LOD} < 2$ e $D' = 1$; rosa representa valores de $\text{LOD} \geq 2$ e $D' < 1$; vermelho representa $\text{LOD} \geq 2$ e $D' = 1$ (LD completo). Pode ser observada a presença de 6 blocos distintos, indicativos de SNPs em forte DL.

Figura 14: Representação do Desequilíbrio de ligação entre os SNPs presentes ao longo da região gênica avaliada através dos valores de D' e LOD score. SNPs definidores dos blocos são destacados em negro.

6.6 ANÁLISE PROMOTORA

Foram identificados 14 SNPs compondo a região promotora do gene *IRF4* com frequência $\geq 1\%$ na amostra populacional estudada (1-17). Os haplótipos foram estabelecidos e ordenados em ordem decrescente de frequência (Tabela 16).

Tabela 16: Pontos de variação presentes ao longo dos haplótipos da região promotora que apresentaram frequência $\geq 1\%$ na amostra populacional estudada

	rs77503657	rs187223943	rs79537043	rs13202461	rs13220460	rs35162881	rs9503555	rs1514346	rs6899334	rs114907235	rs2797305	rs6929421	rs77889581	rs6935510
PROMO001	T	C	G	G	A	A	T	G	T	G	C	C	C	C
PROMO002	T	C	G	G	A	A	T	G	T	G	C	C	C	A
PROMO003	T	C	A	A	A	A	C	A	T	G	T	C	C	C
PROMO004	T	C	G	A	C	G	C	A	C	G	T	C	C	A
PROMO005	T	C	G	G	A	A	T	G	T	G	T	C	C	C
PROMO006	T	C	G	G	A	A	T	G	T	G	T	A	C	A
PROMO007	T	C	G	A	C	G	C	A	C	G	T	C	C	C
PROMO008	T	C	G	A	C	A	T	A	T	G	T	C	T	A
PROMO009	T	C	G	A	C	G	T	A	C	G	T	C	C	A
PROMO010	T	C	G	A	C	A	T	A	C	G	T	C	C	C
PROMO011	T	C	G	G	A	A	T	A	T	G	T	C	C	C
PROMO012	C	T	A	A	A	A	C	A	T	T	T	C	C	C
PROMO013	T	C	G	G	A	A	T	G	T	G	T	C	C	A
PROMO014	C	T	A	A	A	A	C	A	T	G	T	C	C	C
PROMO015	T	C	A	G	A	A	C	A	T	G	T	C	C	C
PROMO016	T	C	G	A	C	A	T	A	C	G	T	C	C	A
PROMO017	C	T	G	A	A	A	C	A	T	G	T	C	C	C

Observação: Os SNPs para os quais foi encontrada associação a algum dos fenótipos de pigmentação estão destacados em negrito. É possível observar que uma única posição distingue os haplótipos PROMO001 e 002 e que as únicas posições variáveis do haplótipo PROMO006 em relação à PROMO002 são os SNPs (rs2797305 e rs6929421).

6.7 ANÁLISE DA REGIÃO CODIFICADORA

Foram identificados 25 SNPs compondo a região codificadora do gene *IRF4* com frequência $\geq 1\%$ na amostra populacional estudada (1-14). Os haplótipos foram estabelecidos e ordenados em ordem decrescente de frequência (Tabela 17).

Tabela 17: Pontos de variação presentes ao longo dos haplótipos da região codificadora que apresentaram frequência $\geq 1\%$ na amostra populacional estudada.

	rs74746812	rs2797308	rs73717071	395857	rs3823305	rs12203592	rs34318727	rs6942173	rs1775589	rs1775590	rs1775591	rs76533660	rs6918132	rs2671422	rs11755347	rs41301853	rs116821462	rs1473037	rs115625962	rs61732392	rs115661683	rs75719951	rs187860014	rs113279656	rs113635483
CD001	G	T	G	C	G	C	G	T	T	T	A	C	G	G	G	C	G	A	G	C	G	C	C	G	A
CD002	G	T	G	C	G	T	G	T	T	T	A	C	G	G	G	C	G	A	G	C	G	C	C	G	A
CD003	G	T	G	C	G	C	G	T	T	T	A	C	G	G	G	C	G	T	G	C	G	C	C	G	A
CD004	G	T	G	C	G	C	G	T	C	C	G	C	G	A	G	T	G	T	G	C	G	C	C	G	A
CD005	G	T	G	C	C	C	G	T	C	C	G	C	G	A	G	C	G	T	G	C	G	G	C	G	A
CD006	G	T	G	C	G	C	A	G	C	C	G	C	G	A	G	C	G	T	G	C	G	C	C	G	A
CD007	G	T	G	C	G	C	A	G	C	C	G	C	C	A	G	C	G	T	G	C	G	C	C	G	A
CD008	G	A	G	C	G	C	G	T	C	C	G	C	G	A	G	C	G	T	G	C	G	C	C	G	A
CD009	G	T	G	C	C	C	G	T	C	C	G	C	G	A	G	C	G	T	G	C	G	C	C	G	A
CD010	G	T	A	C	G	C	G	T	T	T	A	C	G	G	G	C	G	T	G	C	G	C	C	G	A
CD011	G	T	G	C	G	C	G	T	T	T	A	C	G	G	G	C	G	A	G	C	A	C	C	G	A
CD012	G	T	G	C	G	C	A	G	C	C	G	C	G	G	G	C	G	T	G	C	G	C	C	C	A
CD013	G	T	G	C	G	C	G	T	T	T	A	C	G	G	G	C	G	A	G	C	G	C	C	G	G
CD014	G	T	G	C	G	C	G	T	T	C	G	C	G	A	G	C	G	T	G	C	G	C	C	G	A

Observação: Os SNPs para os quais foi encontrada associação a algum dos fenótipos de pigmentação estão destacados em negrito.

6.7.1 ANÁLISE DA REGIÃO *CODING DNA SEQUENCE* (CDS)

Foram identificados 5 SNPs compondo os haplótipos da CDS do gene *IRF4* com frequência $\geq 1\%$ na amostra populacional estudada (1-2). A diferença entre os dois haplótipos foi de somente 1 ponto de variação (rs73717071), para o qual foi encontrada associação a um dos fenótipos de pigmentação (Tabela 18).

Tabela 18: Pontos de variação presentes ao longo dos haplótipos da CDS, que apresentaram frequência $\geq 1\%$ na amostra populacional estudada.

	rs74746812	rs73717071	395922	rs115625962	rs61732392
CDS001	G	G	C	G	C
CDS002	G	A	C	G	C

Observação: O SNP associado a um dos fenótipos de pigmentação avaliados está marcado em negrito. Em destaque o único ponto de variação entre os dois haplótipos da CDS com frequência superior a 1% na população estudada.

6.8 ANÁLISE DE MICRORNAS

Com o uso das ferramentas anteriormente descritas, foi possível estabelecer uma lista dos miRNAs apresentando os 100 maiores valores de ϵ , considerando cada uma das 13 3'UTRs separadamente. Como o interesse do trabalho reside em identificar miRNAs que possuam afinidade diferenciada a sequências polimórficas da 3'UTR, de modo a contribuir para o entendimento das associações observadas, a seleção dos miRNAs de maior interesse foi baseada nos 20 maiores valores de $\Delta\epsilon$ encontrados (Tabela 19).

Tabela 19: microRNAs com alvo para o gene *IRF4* apresentando os 20 maiores valores de $\Delta\epsilon^*$, considerando as UTRs com frequência de pelo menos 1% na amostra estudada

mirna/ UTR média+dp	UTR001 1,2596+0,2127	UTR002 1,2233+0,2183	UTR003 1,2503+0,2222	UTR004 1,2405+0,2122	UTR005 1,2596+0,2127	UTR006 1,2234+0,2179	UTR007 1,2467+0,2069	UTR008 1,2332+0,2106	UTR009 1,3331+0,1833	UTR010 1,2602+0,2121	UTR011 1,2341+0,2201	UTR012 1,3794+0,2309	UTR013 1,2038+0,1907	$\Delta\epsilon$ 0,3174+0,0892
miR-320a	0,9130	0,9195	0,9130	0,9277	0,9130	0,9277	1,2445	0,9195	1,3252	0,9130	0,9130	1,3850	1,2446	0,4720
miR-939-5p	1,4106	0,9416	1,4054	0,9416	1,4106	0,9416	0,9416	0,9416	0,9416	1,4106	1,4054	0,9416	0,9437	0,4690
miR-6754-5p	1,3071	1,3025	1,3072	1,3072	1,3071	1,3025	1,3025	1,3025	1,3664	1,3071	1,3072	0,9336	1,2381	0,4329
miR-1273g-3p	1,3752	1,3681	1,3752	1,3752	1,3752	1,3681	1,3681	1,3681	1,4279	1,3752	1,3752	1,5244	1,1033	0,4211
miR-1273d	1,2848	1,2848	1,2848	1,2899	1,2846	1,2848	1,2848	1,2848	1,6318	1,2848	1,2848	1,6505	1,2630	0,3876
miR-877-5p	1,1605	0,9621	0,9621	1,1667	1,1605	0,9621	1,1667	1,1667	1,3339	1,1605	0,9621	1,3304	1,1667	0,3717
miR-6727-5p	1,5923	1,5876	1,5923	1,5923	1,5923	1,5876	1,5876	1,5876	1,3612	1,5923	1,5923	1,7301	1,5418	0,3689
miR-615-5p	1,2744	1,1778	1,2744	1,2744	1,2744	1,1778	1,1778	1,1778	1,4405	1,2744	1,2744	1,4837	1,1426	0,3411
miR-6786-5p	1,4330	1,4323	1,4330	1,4323	1,4330	1,4323	1,4330	1,4323	1,4323	1,4330	1,1076	1,4330	1,4334	0,3258
miR-6068	1,4551	1,4462	1,4552	1,4552	1,4551	1,4462	1,4462	1,4462	1,4552	1,4551	1,4552	1,4550	1,1361	0,3190
miR-6803-5p	1,0022	1,0022	1,0022	1,0017	1,0022	1,0022	1,0022	1,0022	1,0022	1,0029	1,0022	1,2901	0,9787	0,3114
miR-6775-5p	1,0247	1,0166	1,0166	1,0166	1,0247	1,0166	1,0166	1,0166	1,3030	1,0369	1,0166	1,3196	1,0132	0,3064
miR-1249-5p	1,0424	1,0384	1,0425	1,0342	1,0424	1,0342	1,0425	1,0384	1,3320	1,0424	1,0425	1,2617	1,0340	0,2980
miR-3138	1,0045	1,2754	1,0050	1,2743	1,0045	1,2743	1,0050	1,2754	1,2743	1,0037	1,0050	1,0050	1,0044	0,2717
miR-663a	1,0983	1,0983	1,0983	1,0983	1,0983	1,0983	1,0983	1,0983	1,2992	1,0983	1,0983	1,3429	1,0983	0,2446
miR-6816-5p	1,2153	1,2094	1,2153	1,2153	1,2153	1,2094	1,2094	1,2094	1,3160	1,2153	1,2153	1,4124	1,1844	0,2281
miR-4417	1,3741	1,1805	1,3908	1,1805	1,3741	1,1805	1,3908	1,1805	1,1805	1,3741	1,3908	1,3908	1,4021	0,2216
miR-3141	1,2353	1,2335	1,2353	1,2353	1,2353	1,2335	1,2335	1,2335	1,2403	1,2353	1,2353	1,4232	1,2299	0,1934
miR-6893-5p	1,7871	1,7782	1,7871	1,7871	1,7871	1,7782	1,7782	1,7782	1,7992	1,7871	1,7871	1,8956	1,7129	0,1828
miR-3162-5p	1,2028	1,2109	1,2110	1,2042	1,2028	1,2109	1,2043	1,2042	1,1991	1,2028	1,2110	1,3802	1,2042	0,1811

Observação: : $\Delta\epsilon^*$ - representa a diferença entre os valores de energia de ligação apresentados por cada UTR frente a um microRNA específico. Valores destacados em vermelho são aqueles cujo valor E foi menor que a média apresentada por cada UTR, considerando os 20 microRNAs selecionados.

Foi estabelecida uma lista dos miRNAs apresentando os 100 maiores valores de s , considerando cada uma das 13 3'UTRs separadamente. Como o interesse do trabalho reside em identificar miRNAs que possuam afinidade diferenciada a sequências polimórficas da 3'UTR, a seleção dos miRNAs de maior interesse foi baseada nos 20 maiores valores de Δs encontrados (Tabela 20).

Tabela 20: microRNAs com alvo para o gene *IRF4* apresentando os 20 maiores valores de ΔS^* , considerando as UTRs com frequência de pelo menos 1% na amostra estudada

miRNA/ UTR média+dp	UTR001 2,4113+0,4674	UTR002 2,2930+0,6034	UTR003 2,5053+0,4520	UTR004 2,0507+0,7260	UTR005 2,4108+0,4669	UTR006 2,1801+0,7528	UTR007 2,2839+0,5987	UTR008 2,2643+0,5909	UTR009 2,2434+0,7249	UTR010 2,4113+0,4674	UTR011 2,5060+0,4615	UTR012 2,4397+0,5977	UTR013 2,0660+0,7191	ΔS 1,1759+0,4217
miR-103a-3p	2,7754	2,8004	2,7754	0,7778	2,7754	0,7778	2,7754	2,8004	0,7778	2,7754	2,7754	2,7754	0,7778	2,0226
miR-107	2,8082	2,8351	2,8082	0,8211	2,8082	0,8211	2,8082	2,8351	0,8211	2,8082	2,8082	2,8082	0,8211	2,0140
miR-4672	2,7852	0,8982	2,7905	0,8912	2,7852	0,8912	0,9164	0,8982	0,8912	2,7852	2,7905	0,8500	0,9164	1,9405
miR-3191-3p	2,9431	1,2730	2,9386	1,2730	2,9431	1,2730	1,2730	1,2730	1,2730	2,9431	2,9386	1,2730	1,2730	1,6701
miR-6090	1,3474	1,3278	1,3474	1,3474	1,3474	1,3278	1,3278	1,3278	2,7336	1,3474	1,3474	2,7603	1,2309	1,5295
miR-7705	2,8372	2,8372	2,8372	1,4301	2,8257	2,8372	2,8372	2,8372	2,8372	2,8372	2,8372	2,8372	2,8372	1,4071
miR-6875-5p	2,6776	2,6557	2,6776	2,6776	2,6776	2,6557	2,6557	2,6557	2,8882	2,6776	2,6776	1,7138	2,3154	1,1744
miR-4516	1,6252	1,6252	1,6252	1,6252	1,6252	1,6252	1,6252	1,6252	2,6763	1,6252	1,6252	2,7154	1,6252	1,0902
miR-3660	1,9024	2,8938	2,8938	2,4251	1,9024	2,8938	1,9024	2,2439	2,4251	1,9024	2,8938	1,9024	2,4251	0,9914
miR-1914-3p	2,5446	2,5112	2,5446	2,5446	2,5446	2,5112	2,5112	2,5112	2,7664	2,5446	2,5446	2,9031	1,9153	0,9878
miR-219a-5p	1,9160	2,8451	2,8451	2,8905	1,9160	2,8451	2,8926	2,8905	1,9160	1,9160	2,8451	2,8905	2,8905	0,9766
miR-548p	1,8598	2,8135	1,8598	1,8598	1,8598	2,8135	2,8135	2,8135	1,8598	1,8598	1,8598	1,8598	2,7934	0,9537
miR-7155-3p	2,7797	1,9262	2,7797	2,8751	2,7797	2,8751	2,7797	1,9262	2,8751	2,7797	2,7797	2,7797	2,8751	0,9489
miR-651-3p	1,9919	1,9919	1,9919	1,9919	1,9919	1,9919	1,9919	1,9919	1,9919	1,9919	2,9371	1,9919	1,9919	0,9452
miR-6861-5p	2,8365	2,8328	2,8365	2,8328	2,8365	2,8328	2,8365	2,8328	2,8328	2,8365	1,9060	2,8365	2,8365	0,9305
miR-3065-3p	2,7645	2,7160	2,7645	2,7645	2,7645	2,7160	2,7160	2,7160	2,7945	2,7645	2,7645	2,8416	1,9749	0,8667
miR-660-5p	2,6480	1,9779	2,6480	2,8143	2,6480	2,8143	2,6480	1,9779	2,8143	2,6480	2,6480	2,6480	2,8161	0,8382
miR-653-5p	2,6945	2,6516	2,6516	2,6815	2,6945	2,6516	1,9187	2,6815	2,7470	2,6945	2,6516	2,6815	2,6815	0,8283
miR-3196	2,1459	2,1156	2,1459	2,1459	2,1459	2,1156	2,1156	2,1156	2,5560	2,1459	2,1459	2,8019	2,0021	0,7999
miR-675-5p	2,3438	2,3329	2,3438	2,3438	2,3438	2,3329	2,3329	2,3329	2,3911	2,3438	2,3438	2,9239	2,3206	0,6033

Observação: ΔS^* - representa a diferença entre os valores de especificidade apresentados por cada UTR frente a um microRNA específico. Os valores destacados em vermelho são aqueles cujo valor S foi menor que a média apresentada por cada UTR, considerando os 20 microRNAs selecionados.

Por fim, foi estabelecida uma lista dos miRNAs apresentando os 100 maiores valores de $\Delta\epsilon_s$, considerando cada uma das 13 3'UTRs separadamente. Como o interesse do trabalho reside em identificar miRNAs que possuam afinidade diferenciada a sequências polimórficas da 3'UTR, a seleção dos 20 miRNAs de maior interesse foi baseada nos maiores valores de $\Delta\epsilon_s$ encontrados (Tabela 21).

Tabela 21: microRNAs com alvo para o gene *IRF4* apresentando os 20 maiores valores de $\Delta\epsilon_s^*$, considerando as UTRs com frequência de pelo menos 1% na amostra estudada

miRNA/ UTR média+dp	UTR001 3,8825+1,0701	UTR002 3,6631+1,0138	UTR003 3,8167+1,0944	UTR004 3,7648+0,9988	UTR005 3,8825+1,0701	UTR006 3,6539+1,0192	UTR007 3,8010+1,0345	UTR008 3,7386+0,9943	UTR009 4,1789+0,9994	UTR010 3,8825+1,0700	UTR011 3,7983+1,0605	UTR012 4,0185+1,1187	UTR013 3,6903+0,9688	$\Delta\epsilon$ 1,6425+0,3549
miR-4672	4,5920	1,4349	4,5905	1,4349	4,5920	1,4349	1,4365	1,4349	1,4271	4,5919	4,5905	1,4013	1,5336	3,1907
miR-103a-3p	4,5857	4,6220	4,5764	1,8115	4,5857	1,8113	4,5829	4,6284	1,7990	4,5856	4,5764	4,5075	1,8115	2,8295
miR-107	4,6375	4,6757	4,6281	1,8799	4,6375	1,8797	4,6347	4,6822	1,8672	4,6375	4,6281	4,5585	1,8799	2,8150
miR-3191-3p	5,2418	2,6207	5,2169	2,6210	5,2418	2,6207	2,6210	2,6210	2,6048	5,2418	5,2169	2,5427	2,6210	2,6991
miR-3141	3,5719	2,8937	3,5588	3,5677	3,5719	2,8937	2,9025	2,9026	4,7929	3,5719	3,5588	4,9170	2,7165	2,2006
miR-6090	3,4025	3,3901	3,3899	3,4242	3,4022	3,3901	3,3984	3,3984	5,3807	3,4025	3,3899	5,3456	3,2839	2,0968
miR-6875-5p	4,8073	4,7684	4,7973	4,8044	4,8073	4,7684	4,7750	4,7750	5,1006	4,8073	4,7973	3,0081	4,3303	2,0925
miR-663a	2,6723	2,6395	2,6668	2,6163	2,6723	2,6127	2,6706	2,6431	4,4852	2,6723	2,6668	4,0140	2,6160	1,8724
miR-4516	2,9552	2,9551	2,9549	2,9554	2,9552	2,9551	2,9554	2,9554	4,7896	2,9552	2,9549	4,7813	2,9554	1,8347
miR-18b-3p	4,5335	2,7856	2,7855	2,7908	4,5335	2,7856	2,7908	2,7908	2,7848	4,5335	2,7855	2,7337	2,7982	1,7998
miR-328-5p	4,7778	4,7430	4,7640	4,7735	4,7778	4,7430	4,7520	4,7521	3,5432	4,7778	4,7640	5,3355	4,5220	1,7923
miR-6816-5p	3,5249	2,9001	2,9000	3,5626	3,5249	2,9001	3,5626	3,5626	4,6782	3,5249	2,9000	4,5833	3,5626	1,7782
miR-6742-5p	4,6106	4,5636	4,6030	4,6084	4,6106	4,5636	4,5685	4,5686	4,5405	4,6106	4,6030	2,8401	4,2122	1,7705
miR-1914-3p	4,7858	4,7320	4,7766	4,7831	4,7858	4,7320	4,7380	4,7381	5,0773	4,7857	4,7766	5,2761	3,6500	1,6260
miR-3660	3,0586	4,6120	4,6119	3,8873	3,0586	4,6120	3,0588	3,6630	3,8800	3,0586	4,6119	2,9916	3,8889	1,6204
miR-6861-5p	5,1690	5,1532	5,1576	5,1608	5,1690	5,1532	5,1655	5,1608	5,1487	5,1690	3,6094	5,0666	5,1662	1,5596
miR-6807-3p	4,6479	3,2237	3,2235	4,6174	4,6479	3,2237	4,6174	4,6174	4,6335	4,6479	3,2235	4,5249	4,6174	1,4245
miR-5589-5p	3,2710	3,2612	3,2616	3,2674	3,2710	3,2612	3,2679	3,2674	3,2598	3,2709	4,5531	3,1975	3,2779	1,3556
miR-6870-5p	4,7843	3,5972	4,8656	3,6289	4,7843	3,6193	4,8752	3,6066	3,6279	4,7843	4,8656	4,7890	4,9300	1,3328
miR-3190-5p	3,3526	4,4198	4,4195	4,6846	3,3526	4,6790	4,4247	4,4247	4,6717	3,3526	4,4195	4,2009	4,6846	1,3320

Observação: $\Delta\epsilon_s^*$ - representa a diferença entre os valores combinados de energia/especificidade apresentados por cada UTR frente a um microRNA específico. Valores destacados em vermelho são aqueles cujo valor ϵ_s foi menor que a média apresentada por cada UTR, considerando os 20 microRNAs selecionados.

A seleção de miRNAs realizada a partir do valor Δ (ϵ , s , ϵs) deve ser interpretada com cautela. Isso porque a seleção de um miRNA a partir de uma das pontuações não assegura que este desempenhe um papel biológico. Por exemplo, ao selecionar miRNAs com base apenas no valor s , podemos chegar a uma molécula muito específica para *IRF4*, mas que apresenta um baixo valor (ϵ), o que significa que, mesmo sendo específico, esse miRNA pode não exercer função devido à sua baixa estabilidade. Dessa forma, os 20 miRNAs ordenados de acordo com cada pontuação (ϵ , s , ϵs) foram distribuídos de forma a observar aqueles selecionados simultaneamente por mais de uma pontuação (Tabela 22).

Tabela 22: microRNAs selecionados por apresentarem os 20 maiores valores de ΔE , ΔS e ΔES , distribuídos conforme sua presença em cada um dos índices considerados.

ΔE	ΔS	ΔES
-	miR-103a-3p	miR-103a-3p
-	miR-107	miR-107
miR-1249-5p	-	-
miR-1273d	-	-
miR-1273g-3p	-	-
-	-	miR-18b-3p
-	miR-1914-3p	miR-1914-3p
-	miR-219a-5p	-
-	miR-3065-3p	-
miR-3138	-	-
miR-3141	-	miR-3141
miR-3162-5p	-	-
-	-	miR-3190-5p
-	miR-3191-3p	miR-3191-3p
-	miR-3196	-
miR-320a	-	-
-	-	miR-328-5p
-	miR-3660	miR-3660
miR-4417	-	-
-	miR-4516	miR-4516

Na Tabela 23 são apresentados os 33 pontos de variação que compõem os haplótipos da 3'UTR com frequência $\geq 1\%$ na amostra populacional estudada (1-13).

Tabela 23: Pontos de variação presentes ao longo dos haplótipos da 3'UTR que apresentaram frequência $\geq 1\%$ na amostra populacional estudada

	rs192507789	rs148645823	rs4959248	rs55788549	rs1050975	408035	rs1050976	rs7768807	rs140017346	rs41301859	rs6596980	rs115931740	rs7757384	rs114595725	rs116527520	rs12211228	rs74389255	rs9391997	rs78361121	rs145752540	rs10530949	rs78252705	rs145553686	rs1050979	rs1877175	rs1877176	rs9392502	rs2316514	rs2316515	410865	rs872071	rs45580432	rs2316516
UTR001	C	C	C	C	G	G	T	T	TTGAC	C	A	T	C	C	G	C	G	G	A	G	G	T	C	G	G	G	C	C	G	G	G	A	G
UTR002	T	C	C	C	G	G	C	C	TTGAC	T	A	T	C	C	G	G	G	A	A	G	GTCT	T	C	A	G	G	G	C	A	G	A	A	G
UTR003	T	C	C	C	G	A	C	T	TTGAC	T	A	T	C	C	G	G	G	A	A	G	G	T	C	A	G	G	C	C	G	G	A	A	G
UTR004	C	C	C	C	G	G	C	T	TTGAC	T	A	T	C	C	G	C	G	A	A	G	GTCT	T	C	A	A	A	C	C	A	GT	A	A	T
UTR005	C	C	C	C	G	G	T	T	TTGAC	T	A	T	C	C	A	G	G	G	A	G	G	T	C	G	G	G	C	C	G	G	G	A	G
UTR006	T	C	C	C	G	G	C	C	TTGAC	T	A	T	C	C	G	G	G	A	A	G	GTCT	T	C	A	A	A	C	C	A	GT	A	A	T
UTR007	C	C	C	C	G	A	C	C	TTGAC	T	A	T	C	C	G	G	G	A	A	G	G	T	C	A	G	G	C	C	A	G	A	A	G
UTR008	C	C	C	C	G	G	C	C	TTGAC	T	A	T	C	C	G	G	G	A	A	G	GTCT	T	C	A	G	G	G	C	A	G	A	A	G
UTR009	C	C	C	G	G	A	C	T	TTGAC	T	G	T	C	C	G	G	G	A	A	G	GTCT	T	C	A	A	A	C	C	A	GT	A	A	T
UTR010	C	C	C	C	G	G	T	T	TTGAC	T	A	T	C	C	G	G	G	G	G	G	G	T	C	G	G	G	C	C	G	G	G	A	G
UTR011	T	C	C	C	G	A	C	T	TTGAC	T	A	T	C	C	G	G	G	A	A	A	G	T	C	A	G	G	C	C	G	G	A	T	G
UTR012	C	C	C	C	G	G	C	T	TTGAC	C	G	T	C	C	G	G	G	A	A	G	G	T	C	A	G	G	C	C	A	G	A	T	G
UTR013	C	C	C	C	G	G	C	T	T	T	A	T	C	C	G	G	G	A	A	G	GTCT	T	C	A	A	A	C	C	A	GT	A	A	T

6.9 ANÁLISES DOS DADOS DE ÍNDICE DE MELANINA

Com o objetivo de responder se as classificações realizadas a partir do protocolo de Fitzpatrick e segundo os critérios do IBGE correspondem a uma diferenciação no índice de melanina, comparamos a média de melanina em cada grupo, utilizando para comparação medidas da região mais (testa) e menos exposta (parte interna do braço) ao sol.

Como pode ser observado na Tabela 24, se observou uma média entre cada grupo em torno de 0,40 de melanina, o que confirma esse parâmetro como um bom indicativo para classificação das amostras. Nos tipos de pele mais escuras, a diferença entre os grupos foi maior na medida do braço, evidenciando que peles mais escuras possuem uma maior quantidade do pigmento, mas que quando estimuladas por radiação UV, mesmo uma quantidade menor do pigmento garante um aumento da produção de melanina.

Tabela 24: Comparação das médias do índice de melanina em cada grupo de indivíduos, classificados de acordo com a escala de Fitzpatrick.

Tipo Pele	Média Testa	Média Braço
I+II	1,2397	0,7916
III+IV	1,6312	1,1867
V+VI	2,0737	1,7359

Conforme pode ser observado na Tabela 25, observando as médias dos índices de melanina na testa, é possível perceber que existe uma diferença de 0,25 de melanina diferenciando branco de amarelo; 0,50 diferenciando branco de pardo e 0,80 diferenciando branco de preto. Assim, enquanto amarelo e pardo se diferenciam por 0,25 de melanina, de amarelo para preto é acrescido o dobro, 0,50. Por sua vez, a diferença entre pardo e preto também é de em torno de 0,25 de melanina.

Já observando o braço, a diferença entre os grupos praticamente deixa de existir nos grupos mais claros (branco vs. amarelo). Entre os grupos mais escuros a diferença passa a ser de em torno de 0,40 de melanina.

Tabela 25: Comparação das médias do índice de melanina em cada grupo de indivíduos, classificados de acordo com o IBGE.

IBGE	Média Testa	Média Braço
branco	1,3350	0,9045
amarelo	1,5916	0,9781
pardo	1,8430	1,3771
preto	2,1514	1,8370

A seguir, utilizamos os dados de genótipo referentes aos 18 SNPs associados a alguma das características de pigmentação e a média do índice de melanina em cada genótipo a fim de avaliar se existe uma associação entre determinada variante e a variação no índice de melanina (Tabela 26).

Tabela 26: Análise da associação entre os índices de melanina em cada grupo de genótipos, considerando os 18 SNPs associados a alguma das características de pigmentação avaliadas.

SNP rs	REF	ALT	BRAÇO									TESTA						
			média (desvio padrão)			Mann-Whitney	Kruskal-Wallis	p de R/R x R/A	p de R/R x A/A	p de R/A x A/A	média (desvio padrão)			Mann-Whitney	Kruskal-Wallis	p de R/R x R/A	p de R/R x A/A	p de R/A x A/A
			RR	RA	AA						RR	RA	AA					
						p	p							p	p			
rs2797305	C	T	0,9826 (0,3171)	1,2779 (0,4520)	1,4393 (0,4390)	-	<0,0001	<0,001	<0,001	>0,05	1,4346 (0,3161)	1,5748 (0,4032)	1,8131 (0,3930)	-	<0,0001	<0,001	<0,001	>0,05
rs6929421	C	A	1,1281 (0,4156)	1,2203 (0,4574)	-	0,4419	-	-	-	-	1,5560 (0,3851)	1,6188 (0,3308)	-	0,4928	-	-	-	-
rs77889581	C	T	1,1168 (0,4116)	1,6524 (0,2432)	-	0,0013	-	-	-	-	1,5450 (0,3760)	2,0170 (0,2795)	-	0,0022	-	-	-	-
rs73717071	G	A	1,1168 (0,4116)	1,6524 (0,2432)	-	0,0013	-	-	-	-	1,5450 (0,3760)	2,0170 (0,2795)	-	0,0022	-	-	-	-
rs12203592	C	T	1,1660 (0,4217)	0,8441 (0,2503)	0,8432 (0,1194)	-	0,0006	<0,01	>0,05	>0,05	1,5872 (0,3862)	1,3199 (0,2491)	1,2865 (0,0667)	-	0,0043	<0,05	>0,05	>0,05
rs41301853	C	T	1,1317 (0,4163)	1,1545 (0,4464)	-	0,7617	-	-	-	-	1,5555 (0,3864)	1,6092 (0,3250)	-	0,4787	-	-	-	-
rs75719951	C	G	1,1280 (0,4176)	1,2303 (0,4258)	-	0,3441	-	-	-	-	1,5586 (0,3782)	1,5756 (0,4557)	-	0,9870	-	-	-	-
rs113635483	A	G	1,1270 (0,4110)	1,4197 (0,6456)	-	0,2874	-	-	-	-	1,5557 (0,3820)	1,7299 (0,3685)	-	0,3403	-	-	-	-
rs1131442	G	A	1,1227 (0,4185)	1,1166 (0,4092)	1,1476 (0,4318)	-	0,9316	>0,05	>0,05	>0,05	1,5559 (0,3761)	1,5531 (0,3933)	1,5311 (0,3746)	-	0,9373	>0,05	>0,05	>0,05
rs4959248	T	C	2,2083 (0,0000)	1,3437 (0,4067)	1,1111 (0,4088)	-	0,0138	>0,05	>0,05	>0,05	2,1043 (0,0000)	1,7386 (0,2891)	1,5424 (0,3842)	-	0,0371	>0,05	>0,05	>0,05
rs7768807	T	C	1,1115 (0,4004)	1,1603 (0,4380)	1,1048 (0,4090)	-	0,7103	>0,05	>0,05	>0,05	1,5402 (0,3491)	1,5886 (0,4132)	1,5045 (0,3841)	-	0,5634	>0,05	>0,05	>0,05
rs41301859	T	C	1,1310 (0,4143)	1,1981 (0,5354)	1,0683 (0,0000)	-	0,9742	>0,05	>0,05	>0,05	1,5573 (0,3855)	1,6115 (0,3205)	1,5823 (0,0000)	-	0,8336	>0,05	>0,05	>0,05
rs6596980	G	A	2,2083 (0,0000)	1,2695 (0,4308)	1,1126 (0,4082)	-	0,0463	>0,05	>0,05	>0,05	2,1043 (0,0000)	1,6675 (0,2720)	1,5447 (0,3899)	-	0,085	>0,05	>0,05	>0,05
rs145752540	G	A	1,1342 (0,4223)	1,1034 (0,1845)	-	0,7361	-	-	-	-	1,5644 (0,3831)	1,3785 (0,3038)	-	0,3211	-	-	-	-
rs78252705	T	C	1,1070 (0,3961)	1,8555 (0,3560)	-	<0,0001	-	-	-	-	1,5374 (0,3699)	2,1664 (0,0997)	-	<0,0001	-	-	-	-
rs9392502	C	G	1,1657 (0,4174)	1,0931 (0,4273)	0,9106 (0,2690)	-	0,0889	>0,05	>0,05	>0,05	1,5973 (0,3817)	1,5127 (0,3801)	1,2998 (0,2715)	-	0,0277	>0,05	<0,05	>0,05
rs45580432	A	T	1,1302 (0,4151)	1,1966 (0,4820)	-	0,6633	-	-	-	-	1,5586 (0,3859)	1,5786 (0,3035)	-	0,7543	-	-	-	-
rs140294761	ATG	A	1,1592 (0,4249)	0,9521 (0,3026)	0,6760 (0,0000)	-	0,0359	>0,05	>0,05	>0,05	1,5843 (0,3861)	1,3871 (0,2952)	1,1003 (0,0000)	-	0,0175	<0,05	>0,05	>0,05

Observações: R/R= alelo referência/alelo referência. R/A= alelo referência/alelo alternativo. A/A= alelo alternativo/alelo alternativo. Valores de p menores ou iguais a $\alpha=0,05$ estão destacados em vermelho.

7.1 AMOSTRA POPULACIONAL

A população brasileira é extremamente miscigenada, sendo os critérios utilizados para classificar os grupos de acordo com a cor da pele questionáveis. Em estudo realizado por Pena et al (2011), estes critérios mostraram-se inadequados em representar a ancestralidade da nossa população. Estudos realizados no município de Ribeirão Preto empregando marcadores de ancestralidade evidenciaram que esta população é também altamente miscigenada, sendo os brancos caracterizados por contribuições 79% europeia, 14% africana e 7% ameríndia (FERREIRA et al., 2006), os mulatos por contribuições 62% europeia, 26% africana e 12% ameríndia, enquanto que os negros possuem contribuições 37% europeia e 63% africana (MUNIZ et al., 2008).

Toda essa heterogeneidade se reflete na elevada diversidade fenotípica observada na nossa população, o que por um lado contribui para identificação de novos sítios associados a características de pigmentação (SHRINER et al., 2011), mas por outro, dificulta a classificação quanto à ancestralidade, o que recentemente tem assumido um caráter polêmico devido à atual política de cotas (FRANCIS-TAN, 2016).

A classificação dos indivíduos quanto à ancestralidade é normalmente baseada nos critérios estabelecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e/ou nas características fenotípicas de cada indivíduo (BAILEY; LOVEMAN; MUNIZ, 2013). A fim de avaliar se esses agrupamentos correspondem a variações nos níveis de melanina, foram comparadas medidas do índice de melanina em cada grupo de indivíduos (classificados tanto por meio da escala de Fitzpatrick (FITZPATRICK, 1988) quanto pelos critérios do IBGE), considerando medidas espectrofotométricas realizadas na testa (região normalmente bastante exposta ao sol) e na parte interna do braço (região normalmente pouco exposta ao sol).

Ao observar a divisão dos grupos segundo o critério de Fitzpatrick, percebemos uma variação em torno de 0,40 de melanina nas medidas da testa entre cada grupo (pele clara, média, escura). Nas medidas da parte interna do braço essa mesma variação foi observada entre os grupos mais claros (pele clara e média). No entanto, na comparação entre os tipos de pele média e escura a variação foi maior (0,55). Esse resultado confirma que, a nível basal, a pele escura possui maior quantidade de pigmento e que esta diferença se dilui à medida que nos expomos ao sol, confirmando a problemática de utilizar a cor da pele visível como indicativo de ancestralidade.

Considerando a divisão dos indivíduos de acordo com o critério utilizado pelo IBGE, percebemos uma variação em torno de 0,25 de melanina entre cada grupo (branco-amarelo, amarelo-pardo, pardo-preto), quando considerando as medidas da testa. Ao avaliar as medidas da parte interna do braço percebemos que essa diferença se dilui muito na comparação branco-amarelo (0,07). Isso pode ser explicado pelas discrepâncias observadas entre os tipos de pele branca e asiática. Ainda que asiáticos (leste-asiático) com pele mais clara possuam alguma semelhança em relação à pele branca, estes são mais propensos ao bronzeamento (LOWE; SHAAT; PATHAK, 1997). Essas informações ajudam a explicar porquê da maior diferença (0,25) quando comparamos medidas da testa (devido à maior capacidade de bronzeamento dos asiáticos-descendentes) e porque essa diferença praticamente desaparece (0,07) ao observarmos uma região pouco exposta ao sol.

Na comparação dos tipos de pele mais escura, a diferença entre as medidas de melanina da parte interna do braço torna-se gradativamente maior, o que corrobora a hipótese de que a cor da pele visível não parece ser um bom indicativo de ancestralidade, já que essa diferença se dilui ao observar as medidas de melanina da testa (0,39 para 0,25) na comparação amarelo-pardo e de (0,45 para 0,30) na comparação pardo-preto.

Cabe ainda considerar que outros fatores além dos níveis de melanina contribuem para as diferenças observadas com relação ao tipo de pele. Por exemplo, a derme das peles negra e asiática é mais espessa e compacta que a pele branca, o que ajuda a entender porquê negros e asiáticos têm menos rugas faciais que indivíduos brancos (TAYLOR, 2002).

7.2 DESVIOS EM RELAÇÃO AO EHW

Em relação à amostra total, 2 dos 105 *loci* avaliados apresentaram desvios do EHW. Ao observar as frequências genótípicas de cada um, vemos que os dois loci em questão, 92 (rs1050979) e 101 (rs872071) apresentam pequena deficiência de heterozigotos (aproximadamente 5%), o que leva a um desvio significativo ($p = 0,04978$ e $0,04875$), respectivamente. Considerando os possíveis fatores que podem levar a pequenos desvios como os observados, o mais provável é que esse desvio seja devido a nenhuma outra razão que o acaso. Esse problema de múltiplos testes aparece com

frequência na genômica populacional, sendo observado nestas situações que ao analisar múltiplos SNPs, uma proporção destes deixará de atender ao EHW devido tão somente ao acaso (HARTL; CLARK, 2010). O fato de que a distribuição de haplótipos encontra-se em Equilíbrio de Hardy-Weinberg ($p = 0,8392 \pm 0,0108$) corrobora tal conclusão.

Quando subdividimos a amostra de acordo com as características de pigmentação, encontramos que algumas classes, em determinados *loci* (todos associados a alguma característica fenotípica), se encontram fora do EHW. Para o SNP rs12203592, estão fora do equilíbrio o grupo de pele clara ($p=0,0243$), o grupo apresentando ausência de sardas ($p=0,0016$), olhos mel ($p=0,0434$), 75-100% de cabelos grisalhos ($p=0,0376$) e o grupo representando a amostra total ($p=0,0059$). Para o SNP rs14029471, o grupo com cor de olhos mel ($p=0,0431$) e para o rs2797305 o grupo com cor de cabelo castanho-escuro ($p=0,0383$). Nestes casos, não é esperado o EHW, já que a divisão baseada em grupos fenotípicos rompe com um dos pressupostos do teste.

7.3 ASSOCIAÇÕES ENCONTRADAS

A pigmentação humana é uma característica poligênica e, dessa forma, dependente da regulação da interação entre múltiplos genes (STURM, 2006). Fenótipos de pigmentação tem sido extensivamente estudados, permitindo identificar que a maior parte da variação genética observada entre fenótipos está localizada em regiões não-codificantes do genoma (ALBERT; KRUGLYAK, 2015). Como cerca de apenas 1,5% do genoma humano é constituído por DNA codificante de proteínas (VERNOT et al., 2012), pode-se esperar que uma vasta diversidade de elementos com funções regulatórias no processo de pigmentação venham a ser identificados.

O caso do gene *IRF4*, um fator de transcrição recentemente identificado como associado à pigmentação a partir de um polimorfismo (rs12203592) localizado em uma região intrônica, exemplifica um destes achados (PRAETORIUS et al., 2013a). Considerando que este gene foi recentemente associado à pigmentação e que poucos estudos foram conduzidos neste sentido, o presente trabalho – realizado em uma população bastante miscigenada e utilizando uma metodologia que permite a identificação de variantes pouco frequentes – acrescenta novos conhecimentos quanto a variações presentes não só na região codificante, mas ainda nas regiões regulatórias de *IRF4*.

Neste sentido, as associações encontradas foram divididas de acordo com a região gênica ocupada, sendo observado que dos 18 SNPs associados a algum dos fenótipos de pigmentação avaliados, 3 se encontram na região promotora, 4 em introns, 9 na 3'UTR, 1 na região *downstream IRF4-EXOC2*, e somente 1 se localiza em região de exon.

7.3.1 Associações de SNPs localizados na região promotora

A transcrição de genes eucarióticos é precedida por diversos eventos como remodelamento da cromatina, modificações de histona, ligação de ativadores transcricionais aos *enhancers*, fatores de transcrição aos promotores e recrutamento da RNA polimerase junto ao promotor (SMALE; KADONAGA, 2003). Dessa forma, polimorfismos na região promotora podem afetar o processo de transcrição pela RNA polimerase ao alterar a afinidade entre as proteínas interagindo nesse processo ou recrutando novos fatores de transcrição a se ligarem nessas regiões (ALBERT; KRUGLYAK, 2015). Além disso, variantes em locais de ligação de fatores de transcrição se correlacionam ainda com modificações de histonas, níveis diferenciais de mRNA e metilação do DNA (BANOVICH et al., 2014).

Pode-se dizer que a informação sobre quando e onde um gene será expresso reside em elementos genômicos denominados módulos cis-regulatórios ou elementos cis-regulatórios (YANES-CUNA; KVON; STARK, 2013). Esses módulos cis-regulatórios contêm *clusters* de locais de ligação para fatores de transcrição e moléculas regulatórias, sendo compostos por promotores ou *enhancers* (ISTRAIL; DAVIDSON, 2005).

A transcrição em genes codificantes de proteínas em eucariotos tem início a partir do posicionamento da RNA Polimerase II no local de iniciação correto. Sequências promotoras ficam localizadas na região 5' dos genes, próximo ao sítio de início da transcrição (TSS) e atuam como local de montagem da maquinaria transcricional. O promotor central inclui elementos que podem se estender até em torno de 35pb *upstream* e/ou *downstream* do TSS (SMALE; KADONAGA, 2003).

Enhancers são fragmentos curtos de DNA aos quais se ligam ativadores transcricionais, que por sua vez interagem com os fatores de transcrição, aumentando a afinidade da maquinaria de transcrição a determinados promotores. De modo geral, os

enhancers atuam aumentando a concentração de ativadores próximo ao promotor, aumentando a eficiência e a especificidade com as quais um promotor é reconhecido (LEWIN, 2008). Estes elementos podem ser encontrados em qualquer região do gene, inclusive em introns, motivo pelo qual polimorfismos nessas regiões podem ter efeito ainda que não sejam traduzidos (KNIGHT, 2003). Estes elementos contêm diversos sítios de ligação para ativadores de transcrição específicos e quando colocados fora do local genômico original recobram a ligação ao fator de transcrição endógeno, indicando que a informação cis-regulatória necessária para regulação do padrão de expressão é codificada nas sequências de DNA dos *enhancers* (ISTRIL; DAVIDSON, 2005). É interessante ressaltar que alguns tipos de elementos podem ser encontrados tanto em *enhancer* como em promotores, desempenhando papéis distintos (LEWIN et al., 2011).

A desestabilização do DNA enovelado ao redor do octâmero de histonas é necessária para permitir a passagem da RNA polimerase. Logo, é observado que a estrutura da cromatina em torno dos genes expressos é alterada, sendo essa modificação mais visível na vizinhança de promotores e *enhancers* (FELSENFELD et al., 1996).

Experimentos de ChIP-on-chip e ChIP-seq tem sido utilizados para mapear sítios de ligação de ativadores e fatores de transcrição, bem como modificações de histonas, permitindo a caracterização de regiões regulatórias no genoma (MASTON et al., 2012). O Projeto Encyclopedia of DNA Elements (ENCODE) tem por objetivo fornecer uma lista completa dos elementos funcionais presentes no genoma humano, incluindo elementos que atuam a nível proteico e de RNA e elementos regulatórios que controlam células e momentos nas quais um gene está ativo. Tal projeto fornece ainda diversas informações a respeito de variantes localizadas fora de regiões codificadoras de proteínas, podendo ser acessado diretamente no *UCSC Genome browser* (<http://genome.ucsc.edu/ENCODE/>).

No presente estudo, três SNPs localizados na região promotora foram associados a algum fenótipo de pigmentação: rs2797305 (-2.495pb), rs6929421 (-2.482pb) e rs77889581 (-2.440pb) (Tabela 13). A fim de caracterizar a região onde estão localizados estes SNPs realizamos uma consulta às informações disponíveis na página do projeto ENCODE do no *UCSC Genome browser*, a partir da qual foi possível obter algumas conclusões a respeito da função desempenhada por elementos regulatórios co-localizados com estas variantes (Figura 15).

Promotores e *enhancers* podem apresentar assinaturas epigenéticas distintas quanto à marcação de metilação no quarto resíduo de lisina das histonas H3 (H3K4). A

monometilação em H3K4 foi associada à presença de *enhancers*. Promotores apresentam uma depleção de monometilação da H3K4 no TSS e um enriquecimento dessa modificação mais que 1kb *downstream* ou *upstream*. Em compensação, promotores são enriquecidos pela trimetilação de H3K4 enquanto que em *enhancers* essa modificação costuma não ser substancial (HEINTZMAN et al., 2007). A partir da Figura 15 podemos observar um aumento da trimetilação de H3K4 (H3K4me3), em comparação com a região vizinha ao redor. Essa modificação se relaciona com a atividade transcricional e atua relaxando a estrutura da cromatina, sendo utilizada como um marcador de promotores que estão ativos ou prontos a serem ativados. Ainda é incerto, mas pode ser que a combinação com a presença de H3K4me1 seja indicativo também de *enhancers* e de locais *downstream* do início da transcrição (ROBERTSON et al., 2008).

Por sua vez, a acetilação do resíduo de lisina 27 da histona H3 (H3K27ac) é menos intensa que na região *upstream* aos SNPs em questão. A marcação H3K27ac é associada com iniciação da transcrição e estrutura aberta da cromatina, no entanto, permanece em discussão a possibilidade de que a acetilação promova consequências diferentes de acordo com o resíduo de lisina afetado (HEINTZMAN et al., 2007). A modificação é ainda utilizada para identificação de *enhancers* em atividade (CREYGHTON et al., 2010). Neste caso a modificação H3K27ac observada foi considerada um indicativo da atividade de *enhancer*.

A Dnase I é uma enzima utilizada para mapear a acessibilidade à cromatina. Sítios hipersensíveis à DNase I (DHSs) são marcadores de DNA regulatório e estão relacionados à descoberta de todas as classes de elementos cis-regulatórios, incluindo *enhancers*, promotores, silenciadores, insuladores e regiões de controle do *locus*. Assim, a hipersensibilidade de um sítio à Dnase I, como é observado em cinza na Figura 15, indica uma cromatina relaxada, indicativo de transcrição gênica ativa, uma vez que o remodelamento da cromatina é necessário para ligação dos ativadores de transcrição.

A marcação SYDH TFBS (em azul) indica que esta região consiste em provável sítio de ligação de proteínas regulatórias. No caso dos SNPs avaliados, podemos observar um platô praticamente contínuo em relação a essa marcação, indicando que estes SNPs são alvos para ativadores de transcrição. Experimentos de ChIP-seq, indicaram que a proteína em questão é a PAX5. Tal fator de transcrição foi identificado em (PILZ et al., 1993). Análise complementar realizada no programa HaploReg v4.1 (WARD; KELLIS, 2015) indica que a presença dos alelos C (rs2797305), C

(rs6929421) e C (rs77889581) resultam em alteração de afinidade envolvendo as seguintes proteínas: Myc (rs2797305), BATF, ERalpha-a, Irf e ZBRK1 (rs6929421), e LUN-1 e Nkx2 (rs77889581) e o sítio em questão, resultando em uma alteração da expressão do gene *IRF4*, o que é compatível com as associações observadas. Além disso, muitas dessas proteínas estão expressas em células da pele e do sistema imune, reafirmando os resultados encontrados (<http://www.proteinatlas.org/>).

Ao fim, levando-se em consideração os padrões observados de metilação e acetilação de histonas, a hipersensibilidade à DNase I, a identificação da ligação da proteína PAX5 ao segmento de DNA considerado e distância deste segmento em relação ao TSS, pode-se concluir que a região ocupada por estes SNPs é uma região que influencia a transcrição do gene *IRF4*, atuando como *enhancer*.

Figura 15: Representação gráfica das informações disponibilizadas no ENCODE para a região do gene *IRF4*. Na figura é possível observar as posições ocupadas pelos SNPs rs2797305, rs6929421 e rs77889581 (destacada em amarelo). Na parte superior pode ser observado coordenadas genômicas (hg19). Logo abaixo (em verde) são indicadas as regiões alvo, cobertas pelo ensaio utilizado neste trabalho e a cobertura de reads esperadas após sequenciamento de tal região. Na sequência é possível observar as informações sobre marcas de metilação de histonas (H3K4me1 e H3K4me3), *Transcription Factor Binding Sites* by *ChIP-seq* RNA *Polymerase II* (SYDH TFBS Pol12), H3K27ac, *Digital DNaseI Hypersensitivity Clusters* (Dnase clusters) e *DnaseI hypersensitive sites* (master Dnase IHS).

Até o presente momento, nenhuma associação envolvendo os SNPs rs2797305, rs6929421, rs77889581 foi relatada na literatura. No entanto, estes foram identificados em populações do Projeto 1000 *Genomes*. Assim, podemos observar variação nas frequências alélicas na comparação da população do presente estudo com aquelas avaliadas pelo Projeto 1000 *Genomes* (Tabela 27).

Tabela 27: Frequências alélicas dos SNPs localizados na região promotora do gene *IRF4*, associados a algum dos fenótipos de pigmentação avaliados nas populações que compõem o Projeto 1000 *Genomes* e na população do presente estudo.

SNPs	rs2797305		rs6929421		rs77889581	
pop/alelo	T	C	C	A	C	T
EAS	0,4008	0,5992	1,0000	0,0000	1,0000	0,0000
SAS	0,2413	0,7587	0,9888	0,0112	1,0000	0,0000
AFR	0,6331	0,3669	0,9410	0,0590	0,8941	0,1059
EUR	0,1282	0,8718	0,9901	0,0099	1,0000	0,0000
AMR	0,3127	0,6873	0,9654	0,0346	0,9928	0,0072
Presente estudo	0,2753	0,7247	0,9714	0,0286	0,9846	0,0154

Fonte: 1000 Genomes (<http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>)

No presente estudo, o SNP rs2797305 teve o alelo C e o genótipo CC associados a cabelos loiro-claros (Tabela 13). Conforme é possível observar na Tabela 27, apenas a população AFR apresentou uma frequência maior do alelo T. Por sua vez, a população EUR foi o grupo que apresentou a maior frequência do alelo C, o que é condizente com a associação encontrada, já que indivíduos com cabelos loiros normalmente possuem ancestralidade europeia (GUENTHER et al., 2014). Na presente amostra populacional, as frequências alélicas se encontram mais próximas àquelas encontradas nas populações AMR e SAS. Neste caso, a proximidade em relação à AMR era esperada, uma vez que a população do presente estudo pode ser considerada miscigenada, apresentando certa semelhança ao grupo AMR. O mesmo não é válido para a população SAS, que possui no geral um fenótipo com características mais escuras, principalmente em relação à cor dos cabelos.

Para o SNP rs6929421, o genótipo C/C foi associado a olhos castanho-claros, quando comparados a olhos mel. Ao observar a distribuição de frequências alélicas pode-se observar que a frequência do alelo C foi maior em todas as populações, inclusive na brasileira. Como o menor valor de frequência do alelo C foi encontrado em uma população que apresenta fenótipos mais escuros (AFR), consideramos não haver discrepâncias em relação ao fenótipo associado ao alelo C. A população brasileira apresentou frequências alélicas mais próximas à população AMR, o que era esperado.

O SNP rs77889581 teve o alelo C e o genótipo CC associados a pele clara (Tabela 13). Assim como no caso anterior, a frequência do alelo C foi maior em todas as populações avaliadas, sendo fixado nas populações EAS, EUR e SAS; exceto para a população SAS, este achado seria esperado, uma vez que a característica de pele clara é mais comum em indivíduos das populações EAS e EUR. A população brasileira apresentou frequências alélicas mais próximas à população AMR, o que era esperado.

Com relação às associações envolvendo haplótipos da região promotora (Tabela 14), podemos observar que a PROMO001 foi associada a fenótipos claros de pigmentação (pele clara, olhos azuis, cabelos castanho-escuros na comparação com cabelos pretos, presença de cabelos grisalhos). A PROMO002 foi associada a cabelos loiros, a PROMO003 à presença de cabelos grisalhos e a PROMO006 foi associada a olhos mel (quando comparados a olhos castanho-escuros).

Ao analisarmos as relações observadas entre os haplótipos da promotora identificados com frequência $\geq 1\%$ na nossa população (Tabela 16), verificamos que os haplótipos PROMO001 e PROMO002 diferem em um único nucleotídeo, que não está entre nenhum dos SNPs associados em questão. Como diferença em relação a todos os demais haplótipos considerados, estes são os únicos que apresentam o alelo C no rs2797305. Essa observação nos leva a formular a hipótese de que a presença de rs2797305*C diminuiria a afinidade da ligação de ativadores de transcrição ao *enhancer* em que ele se encontra, não estimulando a expressão de IRF4. Uma vez que o produto de *IRF4* atua na indução da enzima TYR, uma menor expressão do gene levaria à produção de menor quantidade de melanina e, por consequência, de fenótipos mais claros, como foi observado nas associações encontradas envolvendo tanto SNPs individuais, como haplótipos da região promotora. Por fim, concluímos que o rs2797305 se encontra em um *enhancer* que altera a expressão de IRF4 e que o alelo rs2797305*C não estimula a expressão de tal gene.

A PROMO006 se diferencia de todos os demais haplótipos em relação ao alelo rs6929421*A. Esse haplótipo, o único a apresentar o alelo A nessa posição, está associado a um fenótipo de pigmentação mais claro (olhos mel) em comparação com um mais escuro (olhos castanhos-escuros). É interessante observar que tal alelo ocorre apenas em africanos (AFR) e em populações miscigenadas (AMR). A PROMO006 apresenta ainda o alelo rs2797305*T na posição do *enhancer*, o que segundo nossa hipótese levaria a um fenótipo mais escuro. Por sua vez, a presença do alelo rs6929421*A enfraqueceria a ligação de um fator necessário à ativação da transcrição,

diminuindo por outro mecanismo a expressão de IRF4. Assim, indivíduos que possuem o haplótipo em questão apresentam um fenótipo intermediário em relação à cor dos olhos.

Por sua vez, a associação da PROMO003 à presença de cabelos grisalhos não pode ser explicada pelos mecanismos apontados, já que esse haplótipo difere dos demais em relação a pontos que não estão associados a nenhuma característica em questão.

7.1.3 Associação de SNPs encontrados em introns da região coficadora

Polimorfismos genéticos em regiões não-codificantes exercem uma significativa contribuição para variação fenotípica entre os indivíduos (VERNOT et al., 2012). Entretanto, muitas vezes o entendimento sobre a função dos genes na biologia molecular tem sido limitado ao conceito de função proteica (JO; CHOI, 2015). Apenas recentemente essa compreensão passou a ser estendida à regulação da expressão por elementos cis-atuantes, localizados em regiões não-codificantes (CONSORTIUM et al., 2012).

Introns representam cerca de 20-25% do genoma humano (GREGORY, 2005). Estas sequências podem apresentar uma notável conservação e estão na origem do sistema de regulação gênica baseado em RNA (MATTICK, 1994).

Elementos encontrados em introns, como *enhancers* e silenciadores, regulam a iniciação da transcrição pela modulação da função dos promotores de genes (BEAULIEU et al., 2011). O efeito reforçador dos introns na expressão gênica foi conhecido já em 1988, quando demonstrou-se que a remoção destas sequências afeta os níveis de mRNA, levando a uma maior expressão de construções gênicas nas quais as essas sequências foram mantidas (BUCHMANT; BERG, 1988). Além disso, atuam no processo de *splicing* alternativo, permitindo a produção de um grupo de proteínas relacionadas a partir de um único gene (MATTICK, 1994).

O SNP (rs12203592), apresentando a mais forte associação do gene *IRF4* à pigmentação, está localizado em uma região intrônica (HAN et al., 2008). Como já discutido, este SNP atua como um *enhancer*, sendo alvo para o ativador de transcrição TFAP2 α , o qual é por sua vez vizinho a três sítios reconhecidos pelo MITF (Figuras 6 e 7). Juntos, estes fatores de transcrição regulam a expressão de IRF4. A presença do alelo rs12203592*T inibe a ligação de TFAP2 α , e por sua vez de MITF, que neste caso

não estimula a expressão de IRF4. Estando expresso em menor quantidade, a estimulação sobre TYR, resultante da ativação conjunta de MITF-IRF4 deixa de ocorrer, afetando dessa forma a síntese de melanina (PRAETORIUS et al., 2013a).

No presente estudo, o rs12203592*T foi associado a diversas características de pigmentação: cor de pele clara, olhos azuis, presença de sardas e cabelos grisalhos (Tabela 13). Outros estudos já relataram associações entre este polimorfismo e características de pigmentação: cor dos olhos, pele, cabelo e sensibilidade da pele ao sol (HAN et al., 2008), cor dos olhos, cabelo e presença de sardas (ERIKSSON et al., 2010), cor do cabelo, presença de sardas e sensibilidade da pele ao sol (PRAETORIUS et al., 2013a), cor da pele e do cabelo (NORTON et al., 2015), presença de manchas faciais ligadas ao envelhecimento (JACOBS et al., 2015) e presença de cabelos grisalhos (ADHIKARI et al., 2016).

Com relação à cor da pele, o alelo rs12203592*T permitiu diferenciar a pele clara da pele escura com uma probabilidade de associação bastante alta (OR=24,6117; IC95%: 1,4744-410,8283) e um valor de p que permaneceu significativo mesmo após a conservadora correção de Bonferroni ($p=0,0002$). Esse resultado concorda com o achado funcional de que o rs12203592*T levaria a uma expressão reduzida de TYR e, como consequência, a uma diminuição da síntese de melanina (PRAETORIUS et al., 2013a).

Além disso, o genótipo C/C foi associado à pele escura em comparação com pele clara ($p=0,0007$, OR=22,3050, IC95%: 1,3101-379,7429) e o genótipo C/T à pele clara em comparação com pele escura ($p=0,0047$, OR=16,5436, IC95%: 0,9625-284,3647). Por sua vez, o genótipo C/T foi associado no presente estudo a uma redução significativa do índice de melanina em comparação aos grupo com genótipo C/C ($p_{\text{Kruskal Wallis braço}}=0,0006$ e $p_{\text{Kruskal Wallis testa}}=0,0043$), reforçando a associação encontrada do alelo rs12203592*T à pele clara (Tabela 26). A relação do índice de melanina com o SNP rs12203592 já havia sido discutida em um estudo prévio, estando de acordo com o resultado por este encontrado (NORTON et al., 2015).

Interessantemente, o haplótipo da região codificadora mais fortemente associado à pele clara (em comparação com pele escura) neste estudo, CD002 ($p=0,0003$, OR=23,8551, IC95%: 1,4256-399,1698) (Tabela 14), é o único dentre as combinações haplotípicas da região codificadora com frequência acima de 1% nessa população que carrega o alelo rs12203592*T (Tabela 17).

O rs12203592*T foi associado ainda com olhos azuis na comparação com olhos castanho-escuros (OR=11,7500; IC95%: 3,5083-39,3527) com um valor de p que também permaneceu significativo após correção de Bonferroni ($p=4,1829 \times 10^{-8}$). Essa associação corrobora os demais estudos que encontraram associação entre este SNP e cor dos olhos, todos relacionando o rs12203592*T a olhos azuis (HAN et al., 2008; ERIKSSON et al., 2010). O genótipo T/T foi fortemente associado a olhos azuis em comparação com olhos castanho-escuros ($p=0,0413$, OR=20,5319, IC95%: 0,9534-442,1435) e em comparação com olhos castanhos ($p=0,0199$, OR=31,8085, IC95%: 1,4805-683,4054). Por outro lado, o genótipo C/C foi relacionado a olhos castanho-escuros em comparação com olhos azuis ($p=0,0003$, OR=10,8235, IC95%: 2,9292-39,9940) (Tabela 13).

O haplótipo CD002 foi a única combinação haplotípica associada à cor de olhos azuis, cabendo destacar que este haplótipo permitiu diferenciar indivíduos que possuem olhos azuis daqueles que apresentam quaisquer outras colorações de olhos, exceto mel (Tabela 14).

Ainda que com uma associação mais fraca (OR<10), o genótipo C/T foi associado à presença de sardas ($p=0,0312$, OR=3,7255, IC95%: 1,2068-11,5006) no presente estudo. Complementando esse resultado, nenhum haplótipo envolvendo a região codificadora (Tabela 14) ou mesmo qualquer haplótipo formado pelo gene completo (Tabela 15), se mostraram associados à presença de sardas nessa população.

Muito recentemente, o rs12203592*T foi relacionado à predisposição ao aparecimento de cabelos brancos (ADHIKARI et al., 2016). Este é o primeiro estudo, portanto, a confirmar essa associação. É importante destacar que até então nenhum gene havia sido associado a essa característica. Foi encontrada associação do genótipo T/T à classe de indivíduos apresentando 50-75% de cabelos brancos ($p=0,0214$, OR=14,3750, IC95%: 1,4333-144,1721) e 75-100% de cabelos brancos ($p=0,0214$, OR=25,5556, IC95%: 2,1093-309,6160). Corroborando esse achado, o haplótipo CD002, único a apresentar o alelo rs12203592*T, também foi associado à categoria de 75-100% de cabelos brancos ($p=0,0051$, OR=6,6667, IC95%: 2,0183 - 22,0211).

As implicações biológicas dessa associação ainda não estão claras, mas tem-se aventado que o mecanismo de embranquecimento dos fios seja resultado da perda de células não-diferenciadas, fonte de melanócitos no folículo capilar (BUFFOLI et al., 2014). Curiosamente, MITF está envolvido na sobrevivência dos melanócitos via regulação anti-apoptótica de Bcl2, um fator de proteção-chave na resposta do folículo

capilar ao stress oxidativo (WESTGATE; BOTCHKAREVA; TOBIN, 2013). Assim, a hipótese sugerida é de que a menor interação entre *TFAP2α* e o segmento incluindo o alelo rs12203592*T influencie na função desempenhada por MITF na sobrevivência dos melanócitos do folículo capilar (ADHIKARI et al., 2016).

As frequências das variantes presentes na região codificadora de *IRF4* e associadas a algum dos fenótipos de pigmentação avaliados na população do presente estudo e nas populações avaliadas pelo Projeto *1000 Genomes* são apresentadas na Tabela 28.

A frequência alélica muito variável entre as populações mundiais sugere que o alelo associado é alvo de seleção positiva (SULEM et al., 2007). Como pode ser observado em relação ao SNP rs12203592, a frequência do alelo T é maior em europeus, sugerindo que esta variante tenha sido sujeita à seleção positiva nessa população, provavelmente devido a seu efeito na redução da pigmentação da pele (PRAETORIUS et al., 2013a). Tem-se por hipótese que populações vivendo em altas latitudes foram selecionadas a favor de uma pele menos pigmentada, a fim de compensar a menor incidência do sol e facilitar a produção de vitamina D (PARRA, 2007; HUDJASHOV; VILLEMS; KIVISILD, 2013).

Em relação às frequências alélicas encontradas, podemos observar que a frequência do alelo rs12203592*T é mais próxima àquela apresentada pela população AMR, o que está de acordo com o esperado, dado a ancestralidade compartilhada por essas populações. As maiores diferenças na comparação das frequências alélicas de rs12203592*T foram observadas em relação a EAS, onde tal alelo se mostrou ausente, AFR e SAS, populações onde tal alelo não atingiu frequências polimórficas (Tabela 28). É possível que tal alelo tenha surgido no continente africano, onde permaneceu em baixa frequência até o momento em que o homem moderno povoou a Europa. Alternativamente, não se pode descartar a hipótese de que tal alelo tenha surgido na Europa, o que implicaria que sua ocorrência em AFR e SAS fosse decorrente de mistura com europeus.

Tabela 28: Frequências alélicas dos SNPs localizados nos introns da região codificadora do gene *IRF4*, associados a algum dos fenótipos de pigmentação avaliados nas populações que compõem o Projeto 1000 Genomes e na população do presente estudo.

SNPs	rs12203592		rs41301853		rs75719951		rs113635483	
pop/alelo	C	T	C	T	C	G	A	G
AFR	0,9917	0,0083	0,9402	0,0598	0,9992	0,0008	0,9856	0,0144
AMR	0,9280	0,0720	0,9640	0,0360	0,9438	0,0562	0,9885	0,0115
EAS	1,0000	0,0000	1,0000	0,0000	0,8036	0,1964	1,0000	0,0000
EUR	0,8837	0,1163	0,9841	0,0159	1,0000	0,0000	0,9980	0,0020
SAS	0,9939	0,0061	0,9775	0,0225	0,9243	0,0757	1,0000	0,0000
Presente estudo	0,9383	0,0617	0,9626	0,0374	0,9670	0,0330	0,9890	0,0110

Fonte: 1000 Genomes (<http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>)

Além do rs12203592, outros três SNPs intrônicos foram associados a algum dos fenótipos de pigmentação avaliados (rs41301853, rs75719951 e rs113635483). Porém, até então outros estudos não relataram associações a quaisquer dessas três variantes.

O genótipo C/C (rs41301853) foi associado a olhos castanho-escuros em comparação com olhos mel ($p=0,0176$, OR=10,3333, IC95%: 1,8129-58,8981). Complementarmente, o haplótipo CD004 foi associado a olhos mel em comparação com olhos castanho-escuros ($p=0,0247$, OR=8,1429, IC95%: 1,5433 - 42,9648). O haplótipo CD004 é o único a apresentar o alelo T na posição do SNP rs41301853 (Tabela 17). Este achado fortalece a hipótese de que o alelo rs41301853*T esteja associado à coloração de olhos mel na população do presente estudo. Apenas a população AFR apresentou frequência do alelo T maior do que a observada no presente estudo, a população EAS foi monomórfica para este SNP e a menor diferença foi observada em relação à população AMR (Tabela 28).

Para o SNP rs75719951, o alelo C ($p=0,0105$, OR=11,8571, IC95%: 1,4429-97,4399) e o genótipo C/C ($p=0,0117$, OR=11,133, IC95%: 1,3742-90,1964) foram associados a cabelos castanho-escuros em comparação com cabelos pretos. A associação ao mesmo fenótipo foi observada em relação ao haplótipo da região codificadora CD002 ($p=0,0069$, OR=2,0029, IC95%: 1,2310 - 3,2588). Este haplótipo possui, como a imensa maioria dos demais, o alelo rs75719951*C. Apenas o haplótipo CD005, associado a cabelos castanho-claros em comparação com castanho-escuros ($p=0,0449$, OR=8,2703, IC95%: 0,9083 - 75,2998), possui o alelo rs75719951*G. Se apenas o impacto deste SNP na produção de melanina fosse utilizado para explicar essa associação, o resultado encontrado a nível de genótipo e haplotípico seria, em um primeiro momento, contraditório. Contudo, existe a possibilidade de que alelos de outros SNPs presentes neste haplótipo estejam interferindo nas associações encontradas e/ou que o efeito deste polimorfismo não seja direto. De acordo com o observado na

Tabela 28, as populações EAS e SAS foram as que apresentaram a maior frequência do alelo G, enquanto que essa frequência foi muito baixa nas populações de fenótipos de pigmentação mais extremos, AFR e EUR (nula).

O SNP rs113635483 foi associado à presença de cabelos brancos. Pode-se afirmar que esta corresponde à segunda associação genética encontrada até o presente momento para este fenótipo. O alelo rs113635483*A foi associado à classe de indivíduos apresentando acima de 75% de cabelos brancos ($p=0,0204$, $OR=23,1000$, $IC95\%: 2,0063-265,9673$), assim como o genótipo A/G ($p=0,0196$, $OR=25,5556$, $IC95\% 2,1093-309,6160$), também associado a essa categoria. O genótipo A/G foi associado ainda à classe apresentando 50-75% de cabelos brancos ($p=0,0214$, $OR=14,3750$, $IC95\% 1,4333-144,1721$). Com relação às frequências alélicas, pode-se observar que o alelo rs113635483*G não foi encontrado nas populações EAS e SAS e na população EUR apresenta uma frequência muito baixa (Tabela 28). A frequência encontrada na população brasileira é mais próxima àquela observada em AMR e AFR. Porém, essa frequência não é grande o bastante para permitir supor que este alelo esteja sujeito a seleção nessas populações.

Com relação aos haplótipos da região codificadora foram associados à presença de cabelos brancos as combinações haplotípicas: CD002, CD005 e CD013 (Tabela 14). O haplótipo CD002 é o único que possui o alelo rs12203592*T, o qual, como discutido acima, está fortemente associado à classe com 75-100% de cabelos brancos, não explicando dessa forma o efeito do SNP rs113635483. A associação de rs113635483 à classe com 75-100% de cabelos brancos pode ter sido observada, portanto, devido ao DL entre as duas variantes (15).

Por sua vez, o haplótipo CD013 é o único dentre todas as combinações haplotípicas a apresentar o alelo rs113635483*G, tendo sido associado à classe com 50-75% de cabelos brancos ($p=0,0256$, $OR=12,7959$, $IC95\%: 1,3030 - 125,6598$).

O haplótipo CD005 também foi associado a cabelos brancos, porém à classe de 25-50%. Em comparação com os demais haplótipos da codificadora associados ao fenótipo em questão, CD005 apresentou 7 posições com alelos exclusivos e sem nenhum dos alelos associados à característica de cabelos brancos (rs12203592*T e rs113635483*G).

7.1.4 Associação de SNPs encontrados na CDS

A CDS corresponde à sequência de nucleotídeos que darão origem ao mRNA maduro, que será traduzido na ordem sucessiva de aminoácidos na proteína. Uma CDS típica começa com um ATG e termina com um códon de parada. No entanto, a predição da CDS em eucariotos constitui uma tarefa complicada, inicialmente porque não existe um indicador tão bem conservado quanto a sequência de Shine-Dalgarno em procariotos. A sequência de Kozak, uma sequência localizada poucas bases antes do códon de início da tradução AUG, é preditiva para a maioria das CDSs, entretanto não é universalmente conservada em eucariotos (FURUNO et al., 2003). Adicionalmente, a ocorrência de *splicing* alternativo em mais de 75% dos genes eucarióticos complica ainda mais esta atividade (LYNCH, 2004).

No presente trabalho, o local de iniciação da tradução foi estabelecido na posição chr6:393153-393156 (hg19, GRCh37), e o término na posição 407596-407598. Neste intervalo, e desconsiderando as posições em introns, apenas uma associação envolvendo SNPs que codificam aminoácidos foi identificada.

O SNP da CDS associado neste estudo à pigmentação da pele (rs73717071) corresponde a uma substituição *missense* (Ser149Asn). SNPs *missenses* tendem a provocar severas mudanças na função proteica, sendo normalmente sujeitos à seleção negativa (HIRSCHHORN; DALY, 2005). No entanto, a relação entre os efeitos de uma substituição de aminoácido na atividade proteica e sua aptidão evolutiva não é trivial (KRYUKOV; PENNACCHIO; SUNYAEV, 2007). Assim, apesar de SNPs na região codificadora desempenharem uma importante função, seu papel funcional também precisa ser avaliado (HIRSCHHORN; DALY, 2005).

Como muitos dos SNPs para os quais encontrou-se alguma associação, o rs73717071 foi pela primeira vez associado a um fenótipo, neste caso a pigmentação da pele, permitindo diferenciar estatisticamente através do alelo G ($p=0,0101$, OR=19,9565, IC95%: 1,0618-375,0739) e do genótipo G/G ($p=0,0096$, OR=20,9221, IC95%: 1,0994-398,1647) indivíduos de pele clara em comparação aos de pele escura. Complementando este achado, o genótipo G/G apresentou uma redução significativa do índice de melanina em comparação ao grupo com genótipo G/A ($p_{\text{Mann-Whitney}}=0,0013$ e $p_{\text{Mann-Whitney testa}}=0,0022$), reforçando a associação encontrada do alelo rs73717071*G à pele clara (Tabela 26).

Quanto à análise dos haplótipos da CDS, foram encontrados apenas duas combinações com frequência de pelo menos 1% na nossa população (Tabela 18). O haplótipo CD001 foi associado à pele clara em comparação com pele escura ($p=0,0396$, $OR=8,7027$, $IC95\%: 0,9561-79,2147$) e pele clara em comparação com pele média ($p=0,0210$, $OR=8,5740$, $IC95\%: 1,0741-68,4401$), apresentando como única diferença em relação à CD002 a presença do alelo rs73717071*G. Essa observação reforça a possibilidade de que o SNP rs73717071*G esteja realmente associado a uma menor pigmentação da pele nesta população.

Com relação à distribuição mundial das frequências alélicas de rs73717071*G, pode-se verificar que este alelo é fixado nas populações EAS, SAS e EUR, enquanto que nas populações AFR e AMR apresenta frequências de 0,8918 e 0,9927, respectivamente (<http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>). A frequência do alelo rs73717071*G encontrada neste estudo foi de 0,9846, mais próxima àquela observada na população AMR, como observado para vários outros marcadores. Curiosamente, a menor frequência do alelo rs73717071*G foi observada na população AFR, que reconhecidamente possui uma pele mais pigmentada, o que corrobora a associação verificada no presente estudo.

7.1.5 Associação de SNPs encontrados na região 3'UTR

A expressão gênica é em grande parte regulada pela ação de elementos *cis* e *trans* atuantes, como fatores de transcrição e sítios de ligação para essas proteínas. Contudo, nos últimos anos outros elementos, denominados microRNAs, foram identificados como atuando na regulação gênica a nível pós-transcricional e traducional (LEWIS et al., 2003).

MicroRNAs participam da 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). A conservação da região 5' de um miRNA é muito importante para sua ligação aos mRNAs-alvos, especialmente em relação aos nucleotídeos 2-8, que são referidos como região *seed* (LEWIS; BURGE; BARTEL, 2005).

Uma vez que a maior parte dos miRNAs se liga na região 3'UTR (LEWIS et al., 2003) o presente estudo, além de identificar as associações fenotípicas de

polimorfismos presentes na 3'UTR, analisou se os haplótipos dessa região estariam sendo diferencialmente alvejados por miRNAs, devido a diferenças genéticas apresentadas por estas combinações haplotípicas.

A grande maioria dos SNPs associados a fenótipos de pigmentação neste estudo (50% das associações) pertencem à região 3'UTR. Muitos destes não haviam sido até então implicados a qualquer característica de pigmentação ou mesmo a outras condições: rs4959248, rs41301859, rs6596980, rs145752540, rs78252705 e rs45580432.

O genótipo A/A do SNP rs1131442 foi associado no presente estudo a cabelos loiro-claros em comparação com loiro-escuros ($p=0,0274$, OR=13,3333, IC95%: 1,2419-143,1512). O SNP rs1131442 já havia sido associado à leucemia linfocítica crônica (CROWTHER-SWANEOEL et al., 2010a), porém esta é a primeira vez em que este é relacionado a um fenótipo de pigmentação. Essa posição foi uma dentre as desconsideradas ao avaliar as combinações haplotípicas, devido à baixa qualidade dos dados observados, por isso não foi possível avaliarmos esse SNP em um contexto haplotípico. Em relação às frequências alélicas, a menor frequência do alelo rs1131442*G é vista na população AFR, o que é condizente com a associação deste alelo a um fenótipo mais claro de pigmentação. A frequência mais próxima àquela encontrada no presente estudo corresponde à observada na população AMR (Tabela 29).

O SNP rs4959248 permitiu diferenciar indivíduos de pele clara em comparação com pele escura através do alelo C ($p=0,0017$, OR=16,0909, IC95%: 1,9463-133,0279) e do genótipo C/C ($p=0,0044$, OR=14,6667, IC95%: 1,7046-126,1914) como também diferenciou indivíduos de pele clara em comparação com pele média, rs4959248*C ($p=0,0061$, OR=10,6393, IC95%: 1,3594-83,2690), genótipo C/C ($p=0,0054$, OR=11,2558, IC95%: 1,4223-89,0733). Com relação às medidas do índice de melanina, foi possível observar uma variação significativa entre os 3 grupos de genótipo ($p_{\text{Kruskal-Wallis}}=0,0138$) (Tabela 26). Quanto à distribuição populacional das frequências alélicas de rs4959248*C, podemos verificar que este alelo é fixado nas populações EAS e SAS, enquanto que sua menor frequência (0,8260) é observada na população AFR, o que está de acordo com a associação do alelo C a uma pigmentação mais clara da pele.

Tabela 29: Frequências alélicas dos SNPs localizados na região 3'UTR do gene *IRF4*, associados a algum dos fenótipos de pigmentação avaliados nas populações que compõem o Projeto 1000 Genomes e na população do presente estudo.

SNP	REF/ ALT	AFR	AMR	EAS	EUR	SAS	Presente estudo
rs1131442	G	0,6770	0,6167	0,4613	0,6481	0,6074	0,6119
	A	0,3230	0,3833	0,5387	0,3519	0,3926	0,3881
rs4959248	C	0,8260	0,9769	1	0,9970	1	0,9583
	T	0,1740	0,0231	0	0,0030	0	0,0417
rs7768807	T	0,6551	0,7205	0,7272	0,7266	0,6892	0,6930
	C	0,3449	0,2795	0,2728	0,2734	0,3108	0,3070
rs41301859	T	0,9425	0,9726	1	0,9871	0,9908	0,9781
	C	0,0575	0,0274	0	0,0129	0,0092	0,0219
rs6596980	G	0,2050	0,0461	0	0,0159	0,0082	0,0548
	A	0,7950	0,9539	1	0,9841	0,9918	0,9452
rs145752540	G	0,7950	0,9539	1	0,9841	0,9918	0,9868
	A	0,2050	0,0461	0	0,0159	0,0082	0,0132
rs78252705	T	0,8994	0,9914	1	1	1	0,9825
	C	0,1006	0,0086	0	0	0	0,0175
rs9392502	C	0,8805	0,7824	0,7391	0,8121	0,7209	0,8114
	G	0,1195	0,2176	0,2609	0,1879	0,2791	0,1886
rs45580432	A	0,9425	0,9712	1	0,9871	0,9918	0,9759
	T	0,0575	0,0288	0	0,0129	0,0082	0,0241

Fonte: 1000 Genomes (<http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>)

Os haplótipos da 3'UTR: UTR001 e UTR007 foram associados a pele clara em comparação com pele escura e o haplótipo UTR009 a pele clara em comparação com pele média (Tabela 14). Embora todos eles possuam o alelo rs4959248*C, todos os outros haplótipos da região 3'UTR com frequência acima de 1% apresentou o mesmo alelo, fazendo supor que este SNP possa não responder diretamente pelo efeito observado, mas sim pelo DL com outra variação. Quando avaliamos os blocos de DL formado pelas variantes presentes na região 3'UTR do gene *IRF4* podemos observar que o SNP rs4959248 se encontra em forte DL com a variante rs1131442 e em DL completo com diversas outras variantes (Figura 13).

Adicionalmente, destaca-se que a UTR007 possui a presença exclusiva do alelo rs7768807*C. Ao analisarmos os miRNAs associados diferencialmente a esses haplótipos, segundo o score ES (Tabela 19), percebemos que muitos dentre os miRNAs selecionados se ligam diferencialmente de acordo com o haplótipo da 3'UTR considerado (miR-4672, miR-103a-3p, miR-107, miR-3191-3p, miR-3141, miR-6090, miR-6875-5p, miR-663a, miR-4516, miR-18b-3p, miR-328-5p, miR-6816-5p, miR-6870-5p e miR-3190-5p). Assim, pode ser que a associação encontrada reflita um padrão de regulação pós-transcricional desempenhado pela ligação de miRNAs.

O SNP rs7768807 teve o genótipo T/T associado à presença de sardas na população brasileira ($p=0,0422$, $OR=2,7080$, $IC95\%: 1,0596-6,9210$). Este SNP já havia

sido relatado na literatura como associado ao câncer colorretal (SLATTERY et al., 2011) e ao linfoma de Hodgkin (BRODERICK et al., 2010), porém essa é a primeira vez em que é relacionado a um fenótipo de pigmentação. Nenhum haplótipo da 3'UTR ou mesmo do gene completo foi associado à presença de sardas. Em relação à distribuição das frequências alélicas de rs7768807*T nas diferentes populações, observamos a ausência de grande variação geográfica, sendo maior na população EAS e menor na população AFR. Curiosamente, este SNP está em DL com uma outra variante, não associada a qualquer fenótipo de pigmentação neste estudo. Porém, a taxa de recombinação deste bloco com o seguinte é muito baixa, como pode ser percebido pelo valor de D' na Figura 12, indicando que estes possivelmente são transmitidos juntos.

O SNP rs41301859 teve o alelo C associado a olhos mel em comparação com olhos castanho-escuros ($p=0,0332$, OR=17,3636, IC95%: 1,5124-199,3490) e o genótipo T/T associado a olhos castanho-escuros em comparação com olhos mel ($p=0,0321$, OR=19,0000, IC95%: 1,5795-228,5514). Nenhum haplótipo da 3'UTR ou mesmo do gene completo foi associado ao fenótipo em questão. Complementarmente, em relação à frequência alélica de rs41301859*C esta foi maior em AFR e menor em EAS (monomórfico). A frequência de rs41301859*C na nossa população foi mais próxima à AMR, a qual apresentou uma frequência intermediária.

O genótipo A/A do SNP rs6596980 permitiu diferenciar indivíduos de olhos azuis em comparação com olhos mel ($p=0,0261$, OR=19,5263, IC95%: 0,9206-414,1530). A UTR001, que foi associada a olhos azuis em comparação com olhos castanho-escuros ($p=0,0288$, OR=2,2694, IC95%: 1,1075-4,6503) apresenta o alelo rs6596980*A, reafirmando a associação encontrada deste alelo à coloração de olhos azul. Em relação às frequências alélicas nas diferentes populações, a população AFR apresentou a menor frequência do alelo A enquanto que nas demais essa frequência se mostrou mais elevada, o que concorda com a associação encontrada, uma vez que a presença de olhos azuis é frequente em indivíduos de ancestralidade europeia.

O SNP rs145752540 teve o genótipo G/A associado à categoria de 25-50% de cabelos brancos ($p=0,0227$, OR=10,0588, IC95%: 1,5653-64,6375). Dois dos haplótipos da 3'UTR foram associados a cabelos brancos: UTR005 à categoria de 50-75% de cabelos brancos ($p=0,0071$, OR=5,1538, IC95%: 1,6971-15,6515) e UTR011 à categoria de 25-50% de cabelos brancos ($p=0,0235$, OR=9,4545, IC95%: 1,5220-58,7325). A UTR011 foi a única combinação haplotípica a carregar o alelo rs145752540*A, o que permite concluir que o alelo rs145752540*A realmente responda

pela predisposição a fenótipo levemente grisalho (25-50% de cabelos brancos). O alelo rs145752540*A não foi encontrado na população EAS e atingiu maior frequência em AFR (Tabela 29), o que é consistente com o presente achado, uma vez que indivíduos de ancestralidade africana levam mais tempo para se tornarem grisalhos (PANHARD; LOZANO; LOUSSOUARN, 2012). A grande variação nas frequências alélicas de acordo com a população (Tabela 29) pode indicar que este alelo esteja sofrendo algum tipo de seleção (SULEM et al., 2007).

O SNP rs78252705*T ($p=0,0010$, OR=29,5605, IC95%: 1,6451-531,1802) e também o genótipo T/T ($p=0,0008$, OR=31,8767, IC95%: 1,7503-580,5392) se mostraram associados à pele clara em comparação com pele escura. Corroborando essa associação, foi possível observar índice de melanina significativamente menor entre indivíduos com genótipo T/T em comparação com T/C ($p_{\text{Mann-Whitney braço}} = <0,0001$ e $p_{\text{Mann-Whitney testa}} = <0,0001$) (Tabela 26). Apesar de que os haplótipos UTR001 e UTR007 também se mostraram associados a pele clara em comparação com pele escura (Tabela 14), todas as outras combinações haplotípicas com frequência acima de 1% apresentaram o alelo T na posição rs78252705, o que, assim como observado para o SNP rs4959248, faz supor que este SNP possa não responder diretamente pelo efeito observado, mas sim pelo DL com outra variação. O alelo rs78252705*T mostrou-se fixado nas populações da eurásia (Tabela 29) e menos frequente em AFR, o que é consistente com as associações encontradas no presente estudo.

O SNP rs9392502 permitiu diferenciar através do genótipo G/G indivíduos com cabelo loiro-claro de cabelos castanho-escuro ($p=0,0150$, OR=11,5714, IC95%: 1,9576-68,4008), loiro-escuro ($p=0,0168$, OR=24,7333, IC95%: 1,1458-533,9087), e preto ($p=0,0069$, OR=27,0000, IC95%: 2,4639-295,8741), bem como indivíduos com olhos verdes daqueles com olhos castanho-escuros ($p=0,0280$, OR=10,2703, IC95%: 1,1110-94,9388). Quanto às associações de haplótipos da 3'UTR, apenas a UTR002, uma das poucas combinações haplotípicas a apresentar o alelo rs9392502*G, foi relacionada ao fenótipo de cabelos loiro-claros em comparação com pretos ($p=0,0356$, OR=3,4333, IC95%: 1,1205-10,5196). Em compensação, diversas associações à característica de cabelos loiro-claros envolvendo os haplótipos hp007 e hp077 do gene completo foram observadas (Tabela 15). Interessantemente, as populações asiáticas (SAS e EAS), seguidas pela europeia (EUR) apresentam as maiores frequências do alelo G, enquanto AFR possui a menor frequência. A frequência do alelo rs9392502*G no presente estudo

foi mais próxima à encontrada na população EUR. A distribuição da frequência do alelo G é compatível com a associação deste alelo a fenótipos mais claros de pigmentação.

Por fim, quanto ao SNP rs45580432, o alelo T ($p=0,0045$, OR=27,2857, IC95%: 2,7149-274,2269) e o genótipo A/T ($p=0,0040$, OR=31,6667, IC95%: 2,9776-336,7752) foram associados à olhos mel em comparação com olhos castanho-escuros. De modo consistente, o genótipo A/A foi associado a olhos castanho-escuros ($p=0,0040$, OR=31,6667, IC95%: 2,9776-336,7752) e a olhos azuis ($p=0,0261$, OR=19,5263, IC95%: 0,9206-414,1530), em ambos os casos em comparação com olhos mel. Com relação aos haplótipos da 3'UTR, apenas a UTR001 foi associada à coloração dos olhos, no caso olhos azuis, sendo interessante notar que essa combinação haplotípica é uma das que carrega o alelo rs45580432*A. O fato de nenhuma UTR ter se associada a olhos mel não é inesperado, uma vez que apenas duas UTRs pouco frequentes (UTR11 e UTR12) apresentam tal alelo (Tabela 20). Assim como observado quanto ao alelo rs41301859*C, também associado a olhos mel, a frequência do alelo rs45580432*T (Tabela 20) é maior na população AFR e nula na população EAS, o que vai ao encontro da associação encontrada do alelo A em relação a olhos azuis. É importante ressaltar que ambos os marcadores encontram-se em forte DL ($D'=1$, $LOD \geq 2$), e que a UTR11 compartilha tais alelos.

7.1.5.1 MicroRNAs preditos considerando a região 3'UTR de *IRF4*

O gene *IRF4* apresenta uma 3'UTR bastante polimórfica, prova disso é que 50% das associações e 44% dos polimorfismos do gene *IRF4* encontradas neste estudo estão localizadas nessa região do gene. Os sítios de variação da 3'UTR foram organizados em 13 haplótipos com frequência de pelo menos 1% na população do presente estudo (Tabela 23). A seguir, organizamos os resultados obtidos através das três ferramentas de predição *in silico* de interações entre miRNAs e RNA mensageiro, de forma a selecionar miRNAs que se ligam diferencialmente a sequências polimórficas, o que foi realizado ao definirmos a maior variação entre os *scores* apresentados por cada UTR frente à ligação por um miRNA específico (Tabelas 19, 20 e 21). Podemos observar que de maneira geral UTR012 e UTR013 apresentam um comportamento de ligação diferenciado das demais frente a grande parte dos miRNAs.

A abordagem utilizada permitiu a predição de miRNAs que podem se ligar de forma estável a sequências da 3'UTR de *IRF4* (Tabela 19). O valor ϵ reflete a estabilidade da interação miRNA/mRNA. Assim, quanto maior for a pontuação ϵ , maior a probabilidade de ocorrer uma forte interação entre o miRNA e o alvo. Considerando as 3'UTRs associadas a algum fenótipo de pigmentação no presente estudo, isto é, UTR001, UTR002, UTR005, UTR007, UTR009 e UTR011 (Tabela 23), observa-se comportamentos diferenciados frente a miRNAs específicos. Em relação ao miR-877-5p, a UTR009 se ligou com maior intensidade enquanto que UTR002 e UTR011 se ligaram muito fracamente. A UTR011 se diferenciou de todas as demais, por apresentar uma fraca ligação ao miR-6786-5p. Em relação ao miR-6775-5p, a UTR009 se destacou de todas as demais por apresentar uma ligação mais forte. As UTR001, UTR005, UTR007 e UTR011 se ligaram com menor intensidade ao miR-3138 e maior intensidade ao e miR-4417, enquanto que as UTR002 e UTR009 se ligaram com maior e menor intensidade a cada miRNA, respectivamente. A partir destes resultados pode-se observar que existe um padrão semelhante de ligação dos miRNAs entre as UTR001, UTR005, UTR007 e UTR011 e outro entre UTR002 e UTR009. Neste contexto, se destaca a UTR011 por ser alvo de ligações menos intensas de miRNAs e a UTR009 por ser alvo de ligações mais fortes.

A segunda pontuação (denominada valor s) reflete a especificidade da interação miRNA/mRNA. Quanto maior for a pontuação s , maior a probabilidade do miRNA em questão interagir somente com *IRF4* (Tabela 20). Os miR-103a-3p e miR-107 se ligaram menos especificamente à UTR009 e com maior especificidade com UTR001, UTR002, UTR005, UTR007 e UTR011. Em relação aos miR-4672 e miR-3191-3p observou-se que as UTR002, UTR007 e UTR009 são alvos menos específicos do que as UTR001, UTR005 e UTR011. A UTR009 se diferenciou por apresentar uma interação mais específica aos miR-6090 e miR-4516. Através dos resultados observados, podemos observar que a UTR009 possui um comportamento bastante diferente frente às demais UTRs associadas a algum fenótipo de pigmentação.

A terceira pontuação (denominada valor ϵs) foi 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 *IRF4* (Tabela 21). Considerando os miRNAs miR-103a-3p, miR-107, miR-3191-3p e miR-4672, podemos observar um padrão de melhor ligação com as UTR001, UTR005 e UTR011, enquanto que a UTR009 apresentou uma interação muito pior quando

comparada às demais. As UTR002 e UTR007 não são alvos principais de miR-103a-3p e miR-107, embora interajam bem com os dois últimos miRNAs. Em relação aos miR-3141 e miR-6090 somente a UTR009 se liga bem enquanto que em relação à miR-6861-5p somente a UTR011 possui uma ligação prejudicada. Ao final, podemos estabelecer que as UTRs 009 e 011 possuem um padrão diferenciado de ligação aos miRNAs quando comparadas às demais UTRs associadas a algum fenótipo de pigmentação.

Ao observarmos a Tabela 22, percebemos que alguns dos miRNAs que apresentaram uma alta especificidade ao *IRF4*, possuem também uma alta pontuação es, demonstrando que além de específicos, estes miRNAs são estáveis o suficiente para tornarem-se disponíveis na célula e exercer sua função na regulação gênica. Logo, selecionamos como principais candidatos a reguladores da 3'UTR de *IRF4*: miR-103a-3p, miR-3191-3p, miR-3660 e miR-4516. Embora tais candidatos possam ser alvo de estudos funcionais que avaliem a contribuição *in vitro* destas moléculas na regulação da expressão do gene *IRF4* - visando até mesmo a seleção de miRNAs para abordagens terapêuticas na área de dermatologia - a contribuição destes miRNAs no estabelecimento da pigmentação fica condicionada à expressão nos tecidos de interesse, pele, olhos e bulbos capilares.

7.1.5 Associação de SNPs encontrados na região intergênica

Transcritos intergênicos estão submetidos à mesma restrição funcional observada em regiões exônicas, sugerindo que variantes em regiões intergênicas exerçam uma contribuição semelhante àquelas variantes localizadas em regiões de exon, contribuindo dessa forma para diferenças fenotípicas observadas (KHAITOVICH et al., 2006).

O indel rs140294761 está localizado na região entre os genes *IRF4* e *EXOC2* e até então não havia sido associado a qualquer característica de pigmentação ou mesmo a outras condições. O alelo da deleção, rs140294761*A ($p=0,0020$, OR=17,1600, IC95%: 1,0170-289,5494) e o genótipo heterozigoto ATG/A ($p=0,0048$, OR=16,3245, IC95% 0,9498-280,5646) foram associados à pele clara em comparação com pele escura. Por sua vez, o genótipo homozigoto inserção (ATG/ATG) foi associado à pele escura em comparação com pele clara ($p=0,0026$, OR=17,6846, IC95%: 1,0319-303,0636). Por não fazer parte do gene *IRF4*, este SNP não foi incluído em qualquer combinação de

haplótipo gerada. Quanto às medidas do índice de melanina (Tabela 26), foi possível observar uma variação significativa entre os três grupos de genótipo ($p_{\text{Kruskal-Wallis braço}}=0,0359$ e $p_{\text{Kruskal-Wallis testa}}=0,0175$), estando o genótipo ATG/ATG atrelado a um maior índice de melanina, em especial na comparação com o genótipo ATG/A em medidas da testa.

Com relação às frequências do alelo rs140294761*A, observa-se que a menor frequência foi observada na população EAS (0,0010) enquanto que a maior frequência pode ser vista na população do presente estudo (0,0614), valor semelhante ao observado na população EUR (0,0557) (<http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes>). Assim, tanto as medidas do índice de melanina quanto a distribuição das frequências do alelo rs140294761*A nas diferentes populações corroboram a associação deste alelo à pigmentação menos intensa da pele.

8. CONCLUSÕES

Como principais conclusões deste trabalho temos que:

- Dos 18 SNPs associados a algum dos fenótipos de pigmentação no presente estudo, 14 ainda não haviam sido relatados, 1 já havia sido associado a diversas características de pigmentação (rs12203592) e 3 haviam sido associados a outras condições.

- Três SNPs da região promotora foram associados a fenótipos de pigmentação, tendo sido verificado que estes afetam a ligação de fatores de transcrição, podendo levar a diferenças na expressão de *IRF4*.

- Diversas das associações já relatadas envolvendo o SNP rs12203592 a fenótipos de pigmentação foram reafirmadas no presente estudo. Inclusive, este é o primeiro trabalho a confirmar a associação encontrada por Adhikari et al. (2016) do alelo rs12203592*T à presença de cabelos brancos.

- A grande quantidade de polimorfismos encontrados na região 3'UTR (correspondendo a 50% das associações) e a identificação de combinações haplotípicas (UTR009 e UTR011), que de maneira geral apresentam afinidade diferencial aos miRNAs selecionados por apresentar maior especificidade e força de ligação, indicam que este segmento exerce um importante papel no mecanismo de regulação da molécula *IRF4*. Cinco miRNAs foram identificados como principais reguladores em potencial da 3'UTR de *IRF4*: miR-103a-3p, miR-1914-3p, miR-3191-3p, miR-3660 e miR-4516.

- Além das associações de SNPs individuais, foram encontradas diversas associações envolvendo haplótipos nos quais esses SNPs ocorrem. A identificação das combinações haplotípicas permitiu assim reafirmar muitas das associações encontradas a nível de SNP e identificar novas associações a nível haplotípico.

Finalmente, os resultados encontrados reafirmam a importância desempenhada pelo gene *IRF4* na geração da diversidade de fenótipos de pigmentação. Como esperado ao utilizar uma metodologia altamente sensível à detecção de novas variantes e tendo como objeto de estudo uma população bastante miscigenada, diversas novas associações foram identificadas, o que contribui para a seleção de candidatos a estudos funcionais.

9.REFERÊNCIAS BIBLIOGRÁFICAS

9. REFERÊNCIAS BIBLIOGRÁFICAS

ANDREWS, S. FastQC: A quality control tool for high throughput sequence data. Disponível: <http://bioinformatics.babraham.ac.uk/projects/fastqc>. Acesso em: dezembro, 2015.

ADHIKARI, K. et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. **Nature Communications**, p. 1-12, 2016.

ALBERT, F. W.; KRUGLYAK, L. The role of regulatory variation in complex traits and disease. **Nature Reviews Genetics**, v. 16, n. 4, p. 197–212, 2015.

BAILEY, S. R.; LOVEMAN, M.; MUNIZ, J. O. Measures of “Race” and the analysis of racial inequality in Brazil. **Social Science Research**, v. 42, n. 1, p. 106–119, 2013.

BANOVICH, N. E. et al. Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. **PLoS genetics** **PLoS Genetics**, v. 10, n. 9, p. 1–12, 2014.

BARRETT, J. C. et al. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics (Oxford, England)**, v. 21, n. 2, p. 263–5, 15 jan. 2005.

BARTEL, D. P.; LEE, R.; FEINBAUM, R. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. **Cell**, v. 116, p. 281–297, 2004.

BEAULIEU, E. et al. Identification of a novel cell type-specific intronic enhancer of macrophage migration inhibitory factor (MIF) and its regulation by mithramycin. **Clinical and Experimental Immunology**, v. 163, n. 2, p. 178–188, 2011.

BRODERICK, P. et al. IRF4 polymorphism rs872071 and risk of Hodgkin lymphoma. **British Journal of Haematology**, v. 148, n. 3, p. 413–415, 2010.

BUCHMANT, A. R.; BERG, P. Comparison of Intron-Dependent and Intron-Independent Gene Expression. **Molecular and Cellular Biology**, v. 8, n. 10, p. 4395–4405, 1988.

BUDOWLE, B.; VAN DAAL, A. Forensically relevant SNP classes. **BioTechniques**, v. 44, n. 5, p. 603–8, 610, abr. 2008.

BUFFOLI, B. et al. The human hair: from anatomy to physiology. **International Journal of Dermatology**, v. 53, p. 331–341, 2014.

BUSCH, A. et al. IntaRNA : efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. **Bioinformatics (Oxford, England)**, v. 24, n. 24, p. 2849–2856, 2008.

BUTLER, J. M. et al. The development of reduced size STR amplicons as tools for analysis of degraded DNA. **J. Forensic Sci.**, v. 48, n. 5, p-1054-1064, 2003.

CARDON, L. R.; PALMER, L. J. Population stratification and spurious allelic association. **The Lancet**, v. 361, p. 598–604, 2003.

CERQUEIRA, C. C. S. DE et al. Implications of the admixture process in skin color molecular assessment. **PloS one**, v. 9, n. 5, p. e96886, jan. 2014.

CHER, C. Bioanalyzer Applications for Next-Gen. Sequencing: Updates and Tips, Agilent Technologies, 49p, 2011.

CICHOREK, M. et al. Skin melanocytes : biology and development. **Postep Derm Alergol**, p. 30–41, 2013.

CONSORTIUM, E. P. et al. An integrated encyclopedia of DNA elements in the human genome. **Nature**, v. 489, n. 7414, p. 57–74, 2012.

COSTIN, G.; HEARING, V. J. Human skin pigmentation : melanocytes modulate skin color in response to stress. **The FASEB Journal**, v. 21, n. 4, p. 976–994, 2007.

CREYGHTON, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 50, p. 21931–21936, 2010.

CROWTHER-SWANEOEL, D. et al. Fine-scale mapping of the 6p25.3 chronic lymphocytic leukaemia susceptibility locus. **Human Molecular Genetics**, v. 19, n. 9, p. 1840–1845, 2010a.

DA CRUZ, A. T.; JASIULIONIS, M. G. MiRNAs and melanoma: How are they connected? **Dermatology Research and Practice**, v. 2012, p. 1–12, 2012.

DI LEVA, G.; CROCE, C. M. miRNA profiling of cancer. **Current opinion in genetics & development**, v. 23, n. 1, p. 3–11, mar. 2013.

DO, T. N. et al. Biochimica et Biophysica Acta An intronic polymorphism of IRF4 gene in fl uences gene transcription in vitro and shows a risk association with childhood acute lymphoblastic leukemia in males. **Biochimica et biophysica acta**, v. 1802, p. 292–300, 2010.

DUBOIS, P. C. A. et al. UKPMC Funders Group gene expression. **Human Molecular Genetics**, v.23, n. 9, p. 2481–2489, 2010.

DUFFY, D. L. et al. IRF4 variants have age-specific effects on nevus count and predispose to melanoma. **American journal of human genetics**, v. 87, n. 1, p. 6–16, 9 jul. 2010.

DURSO, D. F. et al. Association of Genetic Variants with Self-Assessed Color Categories in Brazilians. **PloS one**, v. 9, n. 1, p. 1–8, 2014.

EDWARDS, A. et al. DNA Typing and Genetic Mapping with Trimeric and Tetrameric Tandem Repeats. n. 2, p. 746–756, 1991.

ERIKSSON, N. et al. Web-based, participant-driven studies yield novel genetic associations for common traits. **PLoS Genetics**, v. 6, n. 6, p. 1–20, 2010.

ESCALANTE, C. R.; AGGARWAL, A. K. Structure of IRF-1 with bound DNA reveals determinants of interferon regulation. **Nature**, v. 391, p. 103–106, 1998.

EWING, B.; HILLIER, L. D.; WENDL, M. C. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. **Genome Research**, p. 175–185, 1998.

EXCOFFIER, L.; LISCHER, H. E. L. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. **Molecular Ecology Resources**, v. 10, n. 3, p. 564–567, 2010.

EXCOFFIER, L.; SLATKIN, M. Incorporating genotypes of relatives into a test of linkage disequilibrium. **American Journal of Human Genetics**, v. 62, n. 1, p. 171–180, 1998.

FELSENFELD, G. et al. Chromatin structure and gene expression. **Proc Natl Acad Sci U S A**, v. 93, p. 9384–9388, 1996.

FERRAGINA, P., MANZINI, G., MAKINEN, V., NAVARRO, G. An Alphabet-Friendly FM-Index. In Proceedings of the 11th International Symposium on String Processing and Information Retrieval. **Lecture Notes in Computer Science**, v. 3246, p. 150–160, 2004.

FERREIRA, B. L. et al. Genomic Ancestry of a Sample Population From the State Report of São Paulo, Brazil. **American journal of human biology**, v. 18, p. 702–705, 2006.

FITZPATRICK, T. B. The validity and practicality of sun-reactive skin types I through VI. **Archives of dermatological research**, v. 124, n. 6, p. 869–871, 1988.

FRANCIS-TAN, A. Light and shadows: An analysis of racial differences between siblings in Brazil. **Social Science Research**, p. 1–12, 2016.

FURUNO, M. et al. CDS Annotation in Full-Length cDNA Sequence CDS Annotation in Full-Length cDNA Sequence. **Genome Research**, v. 13, p. 1478–1487, 2003.

GABRIEL, S. B. et al. The Structure of Haplotype Blocks in the Human Genome. **Science (New York, N.Y.)**, v. 296, p. 2225–2230, 2002.

GIEGERICH, R. et al. Fast and effective prediction of microRNA / target duplexes. **Bioinformatics (Oxford, England)**, v. 10, p. 1507–1517, 2004.

GOODWIN, W. **An Introduction to Forensic Genetics**. [s.l: s.n.].

GREGORY, T. R. Synergy between sequence and size in large-scale genomics. **Nature reviews. Genetics**, v. 6, n. 9, p. 699–708, 2005.

GRIFFITHS-JONES, S. et al. miRBase: microRNA sequences, targets and gene nomenclature. **Nucleic acids research**, v. 34, p. D140–144, 1 jan. 2006.

GUALCO, M, WEISS, ML, BACCHI, M. **Appl Immunohistochem Mol Morphol**, v. 18, n. 4, p. 301–310, 2010.

GUENTHER, C. A. et al. A molecular basis for classic blond hair color in Europeans. **Nat Genet**, v. 46, n. 7, p. 748–752, 2014.

GUPTA, B. S. et al. Interferon Regulatory Factor 4. **J. Exp. Med.**, v. 190, n. 12, p. 1837–1848, 1999.

HAN, J. et al. A Genome-Wide Association Study Identifies Novel Alleles Associated with Hair Color and Skin Pigmentation. **PLoS genetics**, v. 4, n. 5, p. 1–11, 2008.

HALLDORSSON, et al. The Clark Phase-able sample size problem: long range phasing and loss of heterozygosity in GWAS. B. Berger (Ed): p.158-173, Springer-Verlag, Berlin Heidelberg, 2010.

HARADA, H.; TANIGUCHI, T.; TANAKA, N. The role of interferon regulatory factors in the interferon system and cell growth control. **Biochimie**, v. 80, p. 641–650, 1998.

HEINTZMAN, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. **Nature genetics**, v. 39, n. 3, p. 311–8, 2007.

HIRSCHHORN, J. N.; DALY, M. J. Genome-wide association studies for common diseases and complex traits. **Nature reviews Genetics**, v. 6, n. 2, p. 95–108, 2005.

HOLICK, M. F. The Influence of Vitamin D on Bone Health Across the Life Cycle The Vitamin D Epidemic and its Health Consequences 1 – 4. 2005.

HUDJASHOV, G.; VILLEMS, R.; KIVISILD, T. Global patterns of diversity and selection in human tyrosinase gene. **PloS one**, v. 8, n. 9, p. e74307, jan. 2013.

HUNTZINGER, E.; IZAURRALDE, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. **Nature reviews. Genetics**, v. 12, n. 2, p. 99–110, 2011.

ISTRAIL, S.; DAVIDSON, E. H. Gene Regulatory Networks Special Feature: Logic functions of the genomic cis-regulatory code. **Pnas**, v. 102, n. 14, p. 4954–4959, 2005.

ITO, S.; WAKAMATSU, K. Diversity of human hair pigmentation as studied by chemical analysis of eumelanin and pheomelanin. **Journal of the European Academy of Dermatology and Venereology**, v. 25, p. 1369–1380, 2011.

JACOBS, L. C. et al. A Genome-Wide Association Study Identifies the Skin Color Genes IRF4, MC1R, ASIP, and BNC2 Influencing Facial Pigmented Spots. **Journal of Investigative Dermatology**, v. 135, p. 1735–1742, 2015.

JAYANTHY, A.; SETALURI, V. Light-regulated microRNAs. **Photochemistry and Photobiology**, v. 91, n. 1, p. 163–172, 2015.

JEFFREYS, A. J.; WILSON, V. Sir Alec Jeffreys Probe for variable human DNA Isolation of minisatellites Highly polymorphic minisatellites. **Nature**, v. 314, p. 1–5, 1985.

JO, B.; CHOI, S. S. Introns: The Functional Benefits of Introns in Genomes. **Genomics & Informatics**, v. 13, n. 4, p. 112–118, 2015.

JOHN, B. et al. Human MicroRNA Targets. **PLoS biology**, v. 2, n. 11, p. 1862–1879, 2004.

KAYSER, M.; DE KNIJFF, P. Improving human forensics through advances in genetics, genomics and molecular biology. **Nature reviews. Genetics**, v. 12, n. 3, p. 179–92, mar. 2011.

KAYSER, M.; KNIJFF, P. DE. Improving human forensics through advances in genetics , genomics and molecular biology. **Nature Reviews Genetics**, v. 12, p. 179–192, 2011.

KHAITOVICH, P. et al. Functionality of intergenic transcription: An evolutionary comparison. **PLoS Genetics**, v. 2, n. 10, p. 1590–1598, 2006.

KLEIN, U. et al. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. **Nature immunology**, v. 7, n. 7, p. 773–782, 2006.

KNIGHT, J. C. Functional implications of genetic variation in non-coding DNA for disease susceptibility and gene regulation. **Clin. Sci.**, v. 104, n. 5, p. 493–501, 2003.

KOLB, H. Gross Anatomy of the Eye. **WEBVISION**, p. 1–9, 2007.

KRYUKOV, G. V; PENNACCHIO, L. A; SUNYAEV, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. **American journal of human genetics**, v. 80, n. 4, p. 727–739, 2007.

LANGO ALLEN, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. **Nature**, v. 467, n. 7317, p. 832–8, 14 out. 2010.

LE, C. T. **Introductory Biostatistics**. 2nd. ed. New Jersey, USA: John Wiley & Sons, Inc., 2003.

LEITE, T. K. M. et al. Genomic ancestry, self-reported “color” and quantitative measures of skin pigmentation in Brazilian admixed siblings. **PloS one**, v. 6, n. 11, p. e27162, jan. 2011.

LEWIS, B. P. et al. Prediction of Mammalian MicroRNA Targets. **Cell**, v. 115, n. 7, p. 787–798, 2003.

LEWIS, B. P.; BURGE, C. B.; BARTEL, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. **Cell**, v. 120, n. 1, p. 15–20, 14 jan. 2005.

LEWONTIN, R. C.; KOJIMA, K. The Evolutionary dynamics of complex polymorphisms. **Evolution**, v. 14, n. 4, p. 478–472, 1960.

LI, H. et al. The Sequence Alignment / Map format and SAMtools. **Bioinformatics (Oxford, England)**, v. 25, n. 16, p. 2078–2079, 2009.

LIN, J. Y.; FISHER, D. E. Melanocyte biology and skin pigmentation. **Nature**, v. 445, p. 843–850, 2007.

LINS, T. C. et al. Genetic composition of Brazilian population samples based on a set of twenty-eight ancestry informative SNPs. **American journal of human biology : the official journal of the Human Biology Council**, v. 22, n. 2, p. 187–92, 2010.

LIU, F. et al. Digital quantification of human eye color highlights genetic association of three new loci. **PLoS genetics**, v. 6, n. 5, p. e1000934, maio 2010.

LIU, F.; WEN, B.; KAYSER, M. Seminars in Cell & Developmental Biology Colorful DNA polymorphisms in humans. **Seminars in Cell and Developmental Biology**, v. 24, n. 6-7, p. 562–575, 2013.

LYNCH, K. W. Consequences of regulated pre-mRNA splicing in the immune system. **Nature reviews. Immunology**, v. 4, n. 12, p. 931–940, 2004.

MANTA, F. S. N. et al. Analysis of genetic ancestry in the admixed Brazilian population from Rio de Janeiro using 46 autosomal ancestry-informative indel markers. **Annals of human biology**, v. 40, n. 1, p. 94–8, jan. 2013.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual review of genomics and human genetics**, v. 9, p. 387–402, jan. 2008.

MARINN, R. M.; VANNÍCEK, J. Efficient use of accessibility in microRNA target prediction. **Nucleic Acids Research**, v. 39, n. 1, p. 19–29, 2011.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, v. 11, p. 5–7, 2011.

MASTON, G. A. et al. Characterization of Enhancer Function from Genome-Wide Analyses. **Annual review of genomics and human genetics**, v. 13, p. 29–57, 2012.

MATTICK, J. S. Introns: evolution and function. **Current opinion in genetics & development**, v. 831, p. 823–831, 1994.

MCKENNA, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. **Genome research**, v. 20, p. 1297–1303, 2010.

MEDLAND, S. E. et al. Common Variants in the Trichohyalin Gene Are Associated with Straight Hair in Europeans. p. 750–755, 2009.

MIONE, M.; BOSSERHOFF, A. MicroRNAs in melanocyte and melanoma biology. **Pigment Cell and Melanoma Research**, v. 28, n. 3, p. 340–354, 2015.

MILLER, S.A.; DYKES, D.D.; POLESKY, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. **Nucleic Acids Research**, v. 16, 1988.

MITTRUCKER, H. et al. Requirement for the Transcription Factor LSIRF / IRF4 for Mature B and T Lymphocyte Function. **Science (New York, N.Y.)**, v. 275, p. 540–543, 1989.

MUNIZ, Y. C. et al. Genomic ancestry in urban Afro-Brazilians. **Ann Hum Biol**, v. 35, n. 1, p. 104–111, 2008.

NORDBORG, M.; TAVARÉ, S. Linkage disequilibrium: what history has to tell us. **Trends in genetics : TIG**, v. 18, n. 2, p. 83–90, fev. 2002.

NORTON, H. L. et al. Quantitative assessment of skin, hair, and iris variation in a diverse sample of individuals and associated genetic variation. **American Journal of Physical Anthropology**, v. 00, n. August, 2015.

OZATO, K.; TAILOR, P.; KUBOTA, T. The interferon regulatory factor family in host defense: mechanism of action. **The Journal of biological chemistry**, v. 282, n. 28, p. 20065–9, 13 jul. 2007.

PAKSTIS, A. J. et al. Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs. **European journal of human genetics : EJHG**, v. 20, n. 11, p. 1148–54, nov. 2012.

PALMERO, E. I. et al. Mechanisms and role of microRNA deregulation in cancer onset and progression. **Genetics and Molecular Biology**, v. 34, n. 3, p. 363–370, 2011.

PANHARD, S.; LOZANO, I.; LOUSSOUARN, G. Greying of the human hair: A worldwide survey, revisiting the “50” rule of thumb. **British Journal of Dermatology**, v. 167, n. 4, p. 865–873, 2012.

PARRA, E. J. Human Pigmentation Variation : Evolution , Genetic Basis , and Implications for Public Health. v. 105, p. 85–105, 2007.

PARRA, E. J.; KITTLES, R. A; SHRIVER, M. D. Implications of correlations between skin color and genetic ancestry for biomedical research. **Nature genetics**, v. 36, n. 11 Suppl, p. S54–60, nov. 2004.

PENA, S. D. J. et al. The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. **PloS one**, v. 6, n. 2, p. e17063, jan. 2011.

- PILZ, A. J. et al. Mapping of the human homologs of the murine paired-box-containing genes. **Mammalian Genome**, v. 4, n. 2, p. 78–82, 1993.
- PORTO, I. O. P. et al. MicroRNAs targeting the immunomodulatory HLA-G gene: A new survey searching for microRNAs with potential to regulate HLA-G. **Molecular Immunology**, v. 65, n. 2, p. 230–241, 2015.
- PRAETORIUS, C. et al. A Polymorphism in IRF4 Affects Human Pigmentation through a Tyrosinase- Dependent MITF / TFAP2A Pathway. **Cell**, v. 155, n. 5, p. 1022–1033, 2013.
- REES, J. L. Genetics of hair and skin color. **Annu. Rev. Genet.**, v. 37, p. 67–90, 2003.
- ROBERTSON, A. G. et al. Genome-wide relationship between histone H3 lysine 4 mono-and tri-methylation and transcription factor binding. **Genome research**, v. 18, n. 12, p. 1906–1917, 2008.
- ROBINSON, J. et al. Integrative genomics viewer. **Nature Biotechnology**, v. 29, n. 1, p. 24–26, 2011.
- ROSENBERG, N. A. et al. Genetic Structure of Human Populations. v. 298, n. December, p. 2381–2385, 2002.
- ROUZAUD, F. et al. MC1R and the response of melanocytes to ultraviolet radiation. **Mutation research**, v. 571, n. 1-2, p. 133–52, 1 abr. 2005.
- RUMEL, D. Odds Ratio: Algumas considerações. **Revista Saúde Pública**, v. 20, n. 3, p. 3–8, 1986.
- SAIKI, R. K. et al. Enzymatic Amplification of α -Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. **Science (New York, N.Y.)**, v. 230, p. 1350–1354, 1985.
- SANKARARAMAN, S. et al. Estimating Local Ancestry in Admixed Populations. **The American Journal of Human Genetics**, n.82, p. 290–303, 2008.
- SANTOS, R. V. et al. Color, Race, and Genomic Ancestry in Brazil. **Current Anthropology**, v. 50, n. 6, p. 787–819, dez. 2009.
- SHAFFER, A. L. et al. IRF4 addiction in multiple myeloma. **Nature**, v. 454, p. 226–231, 2008.
- SHAFFER, A. L. et al. IRF4 : Immunity . Malignancy ! Therapy ? **Clin Cancer Res**, v. 15, n. 9, p. 2954–2962, 2009.
- SHRINER, D. et al. Mapping of disease-associated variants in admixed populations. **Genome Biology**, v. 12, n. Figure 1, p. 223, 2011.
- SHRIVER, M. D. et al. Skin pigmentation, biogeographical ancestry and admixture mapping. **Human genetics**, v. 112, n. 4, p. 387–99, abr. 2003.

- SLATTERY, M. L. et al. Interferon-signaling pathway: Associations with colon and rectal cancer risk and subsequent survival. **Carcinogenesis**, v. 32, n. 11, p. 1660–1667, 2011.
- SLOMINSKI, A. et al. Melanin Pigmentation in Mammalian Skin and Its Hormonal Regulation. **Physiol Rev**, v. 84, p. 1155–1228, 2004.
- SMALE, S. T.; KADONAGA, J. T. The RNA Polymerase II Core Promoter. **Annual Review of Biochemistry**, v. 72, n. 1, p. 449–479, 2003.
- STEPHENS, M.; DONNELLY, P. Report A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. **Am. J. Hum. Genet**, v. 73, n. 2002, p. 1162–1169, 2003.
- STEPHENS, M.; SMITH, N. J.; DONNELLY, P. A New Statistical Method for Haplotype Reconstruction from Population Data. **American journal of human genetics**, v. 68, p. 978–989, 2001.
- STURM, R. A; DUFFY, D. L. Human pigmentation genes under environmental selection. **Genome biology**, v. 13, n. 9, p. 248, jan. 2012.
- STURM, R. A. A golden age of human pigmentation genetics. **Trends in Genetics**, v. 22, n. 9, p. 464–468, 2006.
- STURM, R. A. Molecular genetics of human pigmentation diversity. **Human Molecular Genetics**, v. 18, n. R1, p. R9–R17, 15 abr. 2009.
- STURM, R. A.; FRUDAKIS, T. N. Eye colour : portals into pigmentation genes and ancestry. **TRENDS in Genetics**, v. 20, n. 8, p. 327–332, 2004.
- SULEM, P. et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. **Nature genetics**, v. 39, n. 12, p. 1443–52, dez. 2007.
- TANG, H. et al. Reconstructing genetic ancestry blocks in admixed individuals. **American journal of human genetics**, v. 79, n. 1, p. 1–12, jul. 2006.
- TAYLOR, S. C. Skin of color: Biology, structure, function, and implications for dermatologic disease. **Journal of the American Academy of Dermatology**, v. 46, n. 2 III, p. 41–62, 2002.
- THORVALDSDOTTIR, H.; JAMES, T.; JILL, P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. **Briefings in bioinformatics**, v. 14, n. 2, p. 178–192, 2012.
- TOBIN, D. J. The cell biology of human hair follicle pigmentation. **Pigment Cell Melanoma Res.**, v. 24, p. 75–88, 2010.
- VERNOT, B. et al. Personal and population genomics of human regulatory variation. **Genome Research**, v. 22, n. 9, p. 1689–1697, 2012.

VIA, M.; GIGNOUX, C.; BURCHARD, E. G. The 1000 Genomes Project: new opportunities for research and social challenges. **Genome medicine**, v. 2, n. 1, p. 3, jan. 2010.

VISSER, M. et al. Genetic variation in regulatory DNA elements: the case of OCA2 transcriptional regulation. **Pigment cell & melanoma research**, v. 27, n. 2, p. 169–77, mar. 2014.

WALSH, S. DNA Phenotyping : Human Pigmentation Traits from Genetic Data. [2011].

WARD, L. D.; KELLIS, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. **Nucleic Acids Research**, v. 44, n. December 2015, p. gkv1340, 2015.

WESTGATE, G. E.; BOTCHKAREVA, N. V.; TOBIN, D. J. The biology of hair diversity. **International Journal of Cosmetic Science**, v. 35, n. 4, p. 329–336, 2013.

XU, W.-D. et al. Targeting IRF4 in autoimmune diseases. **Autoimmunity reviews**, v. 11, n. 12, p. 918–24, out. 2012.

YAMAGUCHI, Y.; BRENNER, M.; HEARING, V. J. Overview : Architecture of the Skin. v. 282, n. 38, p. 27557–27561, 2007.

YANES-CUNA, J. O.; KVON, E. Z.; STARK, A. Deciphering the transcriptional cis-regulatory code. **Cell**, v. 29, n. 1, p. 11–22, 2013.

ZHANG, M. et al. Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. v. 22, n. 14, 2013.

ANEXO 1. 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 2. TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO
PARA ARMAZENAMENTO DE MATERIAL BIOLÓGICO

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

Pesquisador responsável: Prof. Dr. Celso Teixeira Mendes Junior

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

Prezado (a) doador (a),

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

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

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

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

Assinatura do pesquisador responsável

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

Declaro ainda que:

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

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

Ribeirão Preto, ____ de _____ de 20__.

Assinatura do doador