Universidade de São Paulo
Programa Interunidades de Pós-Graduação em Bioinformática

Jaqueline Yu Ting Wang

Determinação pré-natal não invasiva de paternidade
utilizando micro-haplótipos

São Paulo
2017

Jaqueline Yu Ting Wang

Determinação pré-natal não invasiva de paternidade utilizando micro-haplótipos

Dissertação de Mestrado apresentada ao
Programa Interunidades de Pós-Graduação
em Bioinformática da Universidade de São Paulo
como parte dos requisitos exigidos para a
obtenção do título de Mestre em Ciências

Orientador: Prof. Dr. Helder Takashi Imoto Nakaya

Coorientador: Prof. Dr. André Fujita

São Paulo
2017

FICHA CATALOGRÁFICA

W246d Wang, Jaqueline Yu Ting
Determinação pré-natal não invasiva de paternidade utilizando micro-haplótipos /
Jaqueline Yu Ting Wang, [orient.] Helder Takashi Imoto Nakaya, [coorient.] André Fujita.
São Paulo : 2017.
117 p.

Dissertação (Mestrado) - Universidade de São Paulo
Orientador: Prof. Dr. Helder Takashi Imoto Nakaya
Co-Orientador: Prof. Dr. André Fujita

Programa Interunidades de Pós-Graduação em Bioinformática
Área de concentração: Bioinformática

1. Bioinformática. 2. Micro-haplótipos. 3. Polimorfismo de nucleotídeo único. 4. Pré-natal. 5. Não invasivo. 6. Probabilidade de paternidade. I. Nakaya, Helder Takashi Imoto, orient. II. Fujita, André, co-orient. III. Universidade de São Paulo.

CDD - 572.8

Dedico à minha família, por sempre acreditarem em mim!

Agradecimentos

À minha mãe, pelo seu amor incondicional. Por ter muita paciência comigo em todos os momentos de nervosismo e angústia, sempre me apoiando e acreditando na minha capacidade.

Ao meu pai, por seu apoio silencioso, mas sempre presente. Por sempre me dar um puxão de orelha por não ter feito mais matérias de estatística e por ser o doador dos genes que me ajudam a entender a matemática da vida.

Ao meu irmão, que apesar de ser o mais novo, sempre quis ter o papel de irmão mais velho. Por se preocupar demais comigo, e por sempre querer me forçar a melhorar em todos os aspectos.

Aos meus avós, tios, tias, primas e primos, por sempre orarem por mim e desejar o meu sucesso. Por sempre se preocuparem e me animarem durante o período do mestrado.

Ao Professor Helder Nakaya, por me orientar de maneira excepcional durante essa jornada. Por suas incontáveis folhas de esquemas e ideias para a criação dos algoritmos e *pipelines*, por ser sempre animado e me incentivar a buscar novos conhecimentos.

Ao Professor André Fujita, por sua excelente coorientação deste trabalho. Por me acolher em seu laboratório, pelas reuniões semanais que se tornaram conversas divertidas com os mais diversos assuntos possíveis. Por estar sempre me dando apoio e direcionamento nos problemas encontrados ao longo mestrado.

Ao Professor Anatoly Yambartsev, por sua incrível paciência para entender meus problemas estatísticos e embarcar nessa aventura de me direcionar no fabuloso mundo dos números e modelos matemáticos.

Ao Dr. Martin Whittle por nossas conversas a respeito de genética, testes de paternidade, sequenciamento de DNA e muitos outros assuntos. E por transformar uma entrevista

de emprego em uma dissertação de mestrado.

À empresa Genomic Engenharia Molecular, pelo apoio financeiro à minha pesquisa, por fornecer os dados necessários para a criação do modelo e por acreditarem em mim durante todo o trajeto do mestrado.

Às minhas amigas de graduação, Cuia, Coró, e Thaís. Por nossas conversas e momentos de risadas ao longo dos mais de 10 anos que nos conhecemos e formamos essa amizade pra vida. Em especial à Cuia, por ser minha mentora na vida de mestrandia, me guiando pelos caminhos tortuosos, revisando meus textos inúmeras vezes e me apoiando nas apresentações.

Aos meus amigos do PG, por sempre estarem torcendo por mim nessa jornada. Por nossos encontros semanais regados com comida, conversas e risadas e poder contar com o apoio de todos. Em especial à Sinthia, Suli e Bruna, por sempre compartilharmos nossas histórias de vida.

Aos meus amigos do CSBL, por me darem inúmeras ideias, pelos incontáveis lanches da tarde no 19 e por momentos de risadas memoráveis. Em especial ao Thiago, por ser um orientador nato e ser o revisor de texto mais crítico que eu conheço.

Aos meus amigos do Fujita's Team, por assistirem minhas apresentações confusas e me ajudarem a melhorar cada vez mais. Sempre fazendo carinhas confusas e me dando sugestões para tornar minhas explicações mais claras.

Ao Renato Puga, por me ajudar desde o começo a entender os programas e ferramentas que são utilizados para análise de sequenciamento, sempre com muita paciência para guiar uma aluna confusa.

Aos professores Diogo Meyer, Julia Pavan, Maria Vibranovski, Ana Tahira, Maria Rita Passo Bueno e Helena Brentani, por todas as sugestões, correções, indagações e conversas nos exames de qualificação e defesa deste trabalho.

Às secretárias da Bioinformática, Cris e Patrícia, por sempre terem respostas às minhas dúvidas, por estarem dispostas a me ajudar nos problemas encontrados, sempre com muita educação, alegria e carinho.

*“ Ora, a fé é a certeza das coisas que se esperam,
e a prova das coisas que não se vêem.”*

Hebreus 11:1

Resumo

Testes de paternidade geralmente são feitos analisando amostras de DNA do suposto pai, mãe e criança. Para realizar esse exame antes de a criança nascer era preciso recorrer à métodos invasivos, tais como amniocentese e biópsia de vilo corial. Com a descoberta de DNA fetal livre (fcfDNA) no soro e plasma materno, hoje é possível utilizar técnicas que usem esse fcfDNA diminuindo assim os riscos à saúde do feto e da mãe. Testes de paternidade que analisam *Short Tandem Repeats* (STRs) do fcfDNA, embora possíveis, não são confiáveis, pois muitas vezes há degradação do DNA. Por sua vez, *Single Nucleotide Polymorphisms* (SNPs) têm sido demonstrados como bons candidatos para identificação humana e podem ser obtidos de fragmentos pequenos de DNA (ou seja, mesmo com o DNA degradado). No entanto, SNPs possuem um número limitado de alelos diferentes (entre dois e quatro). Micro-haplótipos são segmentos cromossômicos menores do que 200 pb (pares de bases), contendo dois ou mais SNPs que formam pelo menos três haplótipos distintos. Ao utilizá-los como marcadores genéticos, aumentamos o número de possíveis alelos formados a partir dos SNPs. Como o fcfDNA possui um tamanho de aproximadamente 145 pb, isso é suficiente para conter micro-haplótipos que podem ser sequenciados usando tecnologia de Sequenciamento de Nova Geração (NGS). O objetivo desse projeto é determinar a probabilidade de paternidade usando SNPs dentro de micro-haplótipos. Os micro-haplótipos foram escolhidos com base em literatura prévia e as frequências relativas destes foram calculadas com base nos grupos étnicos dos dados do *1000 Genomes*. Dados brutos de sequenciamento de três amostras de DNA são analisados: o suposto pai, a mãe e o plasma materno (mistura de DNA livre da mãe e do feto). Em seguida, desenvolvemos *scripts* para obter e analisar os genótipos do suposto pai e da mãe, para cada um dos micro-haplótipos escolhidos. Combinando informação genotípica, frequências populacionais e frações fetais (plasma), desenvolvemos um método para calcular a probabilidade de paternidade em casos de não exclusão da mesma.

Abstract

Paternity tests are usually done by analyzing DNA samples from the alleged father, the mother, and the child. To perform this exam before the birth, invasive methods such as amniocentesis and chorionic villus sampling are usually necessary. Fortunately, the discovery of fetal cell-free DNA (cfDNA) in maternal plasma and serum, and the development of techniques to analyze this cfDNA have allowed researchers to reduce the health risk for both fetus and mother. Although paternity tests that analyze Short Tandem Repeats (STRs) from cfDNA are possible, they are not reliable because DNA degradation often occurs. Single Nucleotide Polymorphisms (SNPs) have been demonstrated as good candidates for human identification and they can be obtained from small DNA fragments (even from degraded DNA). However, SNPs have a limited number of different alleles (between two and four). Microhaplotypes are chromosomal segments smaller than 200 bp (base pairs) containing two or more SNPs that form at least three distinct haplotypes. By using them as genetic markers, we increased the number of possible alleles formed from the SNPs. Since cfDNA has approximately 145 bp, this is sufficient to contain microhaplotypes that can be sequenced using Next Generation Sequencing (NGS) technology. The aim of this project is to determine the probability of paternity using SNPs within microhaplotypes. Microhaplotypes were chosen based on previous literature review. The haplotype frequencies were calculated based on the ethnic groups from 1000 Genomes database. Raw DNA sequence data from three DNA samples were analyzed: the alleged father, the mother, and the maternal plasma (mixture of mother and cfDNA). Then, we developed scripts to analyse and obtain the genotypes of the alleged father and mother, for each microhaplotype. By combining genotypic information, population frequencies, and fetal fractions (plasma), we developed a method to calculate the probability of paternity in cases of non-exclusion.

Lista de Figuras

1.1	Exemplo da localização do micro-haplótipo no cromossomo	26
3.1	Ferramenta <i>data slicer</i> do banco de dados <i>1000 Genomes</i>	36
3.2	Gráfico do PC1 <i>versus</i> o PC2 dos SNPs analisados.	39
4.1	<i>Workflow</i> do processamento e análise dos dados.	43
4.2	<i>Workflow</i> da etapa da <i>Qualidade</i>	44
4.3	Extração dos SNPs dos <i>reads</i>	44
4.4	Qualidade das bases e determinação de haplótipos.	45
4.5	Pareamento dos <i>reads</i> em haplótipos.	45
4.6	Simulação utilizando os dados do <i>1000 Genomes</i>	47
4.7	<i>Boxplot</i> da simulação utilizando os dados <i>1000 Genomes</i>	48
4.8	<i>Dotplot</i> do número de falso positivo <i>versus</i> o número de micro-haplótipos.	49
4.9	Histograma da frequência relativa dos haplótipos.	53
5.1	Histograma da frequência relativa dos possíveis haplótipos fetais.	66
5.2	Novo histograma da frequência relativa dos possíveis haplótipos fetais.	67
5.3	<i>Barplot</i> da frequência relativa dos possíveis haplótipos herdados do pai.	68
5.4	Histograma da distância de Hamming dos erros de sequenciamento.	69
5.5	<i>Dotplot</i> dos valores estimado e simulados de γ e δ	73
5.6	<i>Dotplot</i> de P_C^E dos dados da Genomic.	74
5.7	<i>Dotplot</i> de W dos dados da Genomic.	75
5.8	<i>Heatmap</i> da simulação suposto pai <i>versus</i> mãe e plasma, mínimo de 12.	76

5.9	<i>Heatmap</i> da simulação suposto pai versus mãe e plasma, mínimo de 15. . .	77
5.10	<i>Boxplot</i> da simulação da fração fetal.	79
5.11	<i>Barplot</i> da frequência relativa dos possíveis haplótipos herdados do pai na simulação.	81
5.12	<i>Dotplot</i> dos valores estimado e simulados de γ e δ dos dados artificiais. . .	82
5.13	<i>Dotplot</i> de P_C^E dos dados artificiais.	83
5.14	<i>Dotplot</i> de W dos dados artificiais.	84
5.15	<i>Heatmap</i> da simulação suposto pai versus mãe e plasma artificial, mínimo de 12.	85
5.16	<i>Heatmap</i> da simulação suposto pai versus mãe e plasma artificial, mínimo de 15.	86

Lista de Tabelas

1.1	Exemplo da heterozigosidade de um micro-haplótipo.	27
1.2	Comparativo das plataformas Illumina MiSeq e do Ion Torrent PGM	28
1.3	Exemplo de dados em fase do <i>1000 Genomes</i>	29
1.4	Exemplo de dados fora de fase	30
3.1	Lista dos 20 micro-haplótipos escolhidos.	34
3.2	Disposição das informações do <i>1000 Genomes</i>	37
3.3	Lista dos haplótipos do micro-haplótipo M01.	37
4.1	Informações armazenadas nos arquivos BAM e SAM.	42
4.2	Regras para determinação do genótipo.	47
4.3	Lista de possíveis haplótipos do M02 da amostra T58S02.	50
4.4	Lista de possíveis haplótipos do M03 da amostra T58S02.	51
4.5	Lista de possíveis haplótipos do M18 da amostra T58S02.	51
4.6	Lista de possíveis haplótipos do M15 da amostra T58S02.	52
4.7	Lista de possíveis haplótipos do M15 da amostra T48S01.	52
4.8	Lista de possíveis haplótipos do M13 da amostra T48S01.	52
4.9	Novas regras para determinação do genótipo.	55
4.10	Número de micro-haplótipos cobertos em cada amostra.	55
5.1	Regras das evidências de paternidade.	59
5.2	Valores do IP_C e seus significados.	64
5.3	Novas regras das evidências de paternidade.	70

5.4	Lista de qualidade dos trios.	70
5.5	Resultados inconsistentes dos trios.	72
5.6	Número de micro-haplótipos cobertos em cada amostra da simulação. . . .	80
A.1	Populações do <i>1000 Genomes</i>	101
B.1	Lista dos haplótipos do micro-haplótipo M02.	103
B.2	Lista dos haplótipos do micro-haplótipo M03.	103
B.3	Lista dos haplótipos do micro-haplótipo M04.	104
B.4	Lista dos haplótipos do micro-haplótipo M05.	105
B.5	Lista dos haplótipos do micro-haplótipo M06.	106
B.6	Lista dos haplótipos do micro-haplótipo M07.	106
B.7	Lista dos haplótipos do micro-haplótipo M08.	107
B.8	Lista dos haplótipos do micro-haplótipo M09.	107
B.9	Lista dos haplótipos do micro-haplótipo M10.	108
B.10	Lista dos haplótipos do micro-haplótipo M11.	108
B.11	Lista dos haplótipos do micro-haplótipo M12.	110
B.12	Lista dos haplótipos do micro-haplótipo M13.	110
B.13	Lista dos haplótipos do micro-haplótipo M14.	111
B.14	Lista dos haplótipos do micro-haplótipo M15.	111
B.15	Lista dos haplótipos do micro-haplótipo M16.	112
B.16	Lista dos haplótipos do micro-haplótipo M17.	113
B.17	Lista dos haplótipos do micro-haplótipo M18.	113
B.18	Lista dos haplótipos do micro-haplótipo M19.	114
B.19	Lista dos haplótipos do micro-haplótipo M20.	114
C.1	Valores de IP_m para diferentes combinações de haplótipos	117

Lista de Abreviaturas

fcfDNA	DNA fetal livre (<i>fetal cell-free DNA</i>)
ff	Fração fetal
pb	Pares de base
SNP	<i>Single Nucleotide Polymorphism</i>
STR	<i>Short Tandem Repeat</i>
NGS	Sequenciamento de Nova Geração (<i>Next Generation Sequencing</i>)
SP	Suposto pai
M	Mãe
PL	Plasma
AFR	Africana
AMR	Americana
EAS	Leste Asiática
EUR	Européia
SAS	Sul Asiática
ID	Identificador
REF	Alelo referência
ALT	Alelo alternativo
GT	Genótipo
PCA	Análise de Componente Principal (<i>Principal Component Analysis</i>)

Sumário

Introdução	23
1.1 DNA fetal livre	23
1.2 Testes de paternidade	24
1.3 Micro-haplótipos	26
1.4 Sequenciamento de Nova Geração	27
1.5 Dados do projeto <i>1000 Genomes</i>	29
Objetivo	31
2.1 Objetivo principal	31
2.2 Objetivos específicos	31
Definição dos micro-haplótipos	33
3.1 Materiais e Métodos	33
3.1.1 Micro-haplótipos	33
3.1.2 Análise do projeto <i>1000 Genomes</i>	35
3.1.3 <i>Pipeline</i>	36
3.2 Resultados e Discussão	37
Análise dos dados de sequenciamento	41
4.1 Materiais e Métodos	41
4.1.1 Genomic Engenharia Molecular	41

4.1.2	Bioinformática	42
4.2	Resultados e Discussão	47
4.2.1	Simulação dos dados do <i>1000 Genomes</i>	47
4.2.2	Análise dos dados da Genomic Engenharia Molecular	50
Cálculo da probabilidade de paternidade		57
5.1	Materiais e Métodos	57
5.1.1	Evidências de paternidade	57
5.1.2	Probabilidade baseada nas evidências de paternidade	60
5.1.3	Probabilidade de paternidade clássica	61
5.2	Resultados e Discussão	65
5.2.1	Obtenção da fração fetal	65
5.2.2	Evidências de paternidade	68
5.2.3	Probabilidade de paternidade - Evidências de paternidade	73
5.2.4	Probabilidade de paternidade - Clássica	75
5.2.5	Validação do método em dados simulados	78
Conclusão		89
6.1	Considerações finais	89
6.2	Trabalhos futuros	93
Referências Bibliográficas		94
A Populações do <i>1000 Genomes</i>		101
B Haplótipos do micro-haplótipo		103
C Índice de Paternidade		117

Introdução

No presente trabalho, nosso principal objetivo é desenvolver um teste de paternidade para ser realizado antes do nascimento da criança. A obtenção da amostra fetal é através de sangue materno apenas, evitando métodos que possam trazer riscos à gravidez. Nesse capítulo, apresentamos uma introdução aos conceitos que serão utilizados durante o trabalho, técnicas já existentes e suas limitações e o banco de dados público de onde será obtida a informação populacional. Primeiramente, apresentamos a motivação, o objetivo principal e os objetivos secundários.

Posteriormente, no Capítulo 3, apresentamos os marcadores genéticos utilizados (micro-haplótipo), como estas regiões foram escolhidas e a análise e obtenção das frequências populacionais a partir do banco de dados do projeto *1000 Genomes*. No Capítulo 4 apresentamos a obtenção dos dados de sequenciamento das amostras de suposto pai, mãe e plasma e o método de processamento e análise dos mesmos para obter as informações genéticas contidas nos três tipos de amostras. Seguindo para o Capítulo 5, onde utilizamos as frequências populacionais obtidas para cada micro-haplótipo, os dados genéticos do suposto pai e da mãe, e as informações do plasma e fração fetal para calcular a probabilidade de paternidade nos casos onde não há a exclusão.

1.1 DNA fetal livre

Em 1989, um estudo realizado por Lo e colaboradores foi capaz de detectar células fetais no sangue de gestantes, tornando-o uma alternativa mais segura para a realização de exames diagnósticos no feto [1]. Posteriormente, foi detectada também a presença de DNA fetal livre (*fetal cell-free DNA*, fcfDNA) no plasma e soro materno, juntamente com DNA livre da mãe [2]. Com isso, o desenvolvimento de técnicas para analisar o fcfDNA

se tornou muito útil, dada sua maior concentração em relação às células fetais no mesmo volume de sangue [2]. Durante a gestação, a fração fetal (ff; fcfDNA/DNA livre total) desse DNA circulante aumenta progressivamente, atingindo aproximadamente 10% entre a 11^a e a 13^a semana de gestação [3]. A partir da 10^a até a 21^a semana, aumenta 0,1% semanalmente, e após esse período, o aumento é de 1% por semana [4]. Contudo, a ff na décima semana tem desvio padrão de aproximadamente 5% [5] e aproximadamente 2% das gestantes terão uma ff < 4% [4]. Além disso, existe um decréscimo da ff com o aumento do peso materno, provavelmente isso se deve a efeitos de diluição [3, 4].

Desde a sua descoberta, o fcfDNA é utilizado em muitos estudos de investigação fetal e atualmente é possível obter algumas informações genética do feto, tais como desordens relacionadas ao sexo, aneuploidias, desordens autossômicas recessivas e inclusive investigação da paternidade [6–9]. Para se ter acesso ao fcfDNA, uma amostra do sangue periférico da gestante deve ser obtido, misturado com anticoagulante e centrifugado, de forma a se obter o plasma. Este contém uma mistura de DNA genômico fragmentado materno e fetal, com tamanho médio de 166 pb e 145 pb respectivamente, provavelmente resultante de processos apoptóticos normais de ambos indivíduos [2, 10].

1.2 Testes de paternidade

O teste de paternidade envolve analisar marcadores genéticos de amostras de DNA do suposto pai, mãe e criança. Geralmente é realizado utilizando amostras de sangue dos indivíduos, entretanto em algumas situações há o desejo ou a necessidade de realizar o teste antes do nascimento. Para se realizar este tipo de exame investigativo, procedimentos invasivos tais como amniocentese e biópsia de vilo corial são então realizados. Porém, estes procedimentos oferecem riscos à gravidez e podem acarretar um aborto [11]. Uma possível alternativa, para se realizar os testes de vínculo genético de filiação pré-natal não invasivo, é analisar marcadores genéticos no DNA genômico da mãe, do suposto pai e do plasma da gestante (mistura de DNA livre da mãe e do feto).

Tradicionalmente, em testes de paternidade, os marcadores genéticos utilizados são os *Short Tandem Repeats* (STRs). STRs são regiões do DNA altamente polimórficas compostas por um conjunto de sequências (2-7 pb) que se repetem de 5 a 30 vezes em uma

determinada localização do genoma [12–15]. Analisando um conjunto de 13 a 15 *loci* de STRs, é possível diferenciar membros de uma população, testar paternidade, resolver casos de pessoas desaparecidas, crimes violentos e desastres envolvendo um grande número de pessoas [12, 16, 17]. Porém, o grande tamanho dos *amplicons* de PCR faz com que amostras de DNA muito degradado sejam difíceis de analisar, e a interpretação de misturas pode ser complicada pela presença de *stutter*, um artefato do processo de PCR que pode gerar um falso alelo menor ou maior do que o alelo principal [12, 18, 19]. Para contornar o problema de amostras de DNA degradado, o uso de miniSTRs se tornou uma alternativa muito viável, por gerar *amplicons* menores [13]. No entanto, misturas de DNA ainda não foram totalmente resolvidas para este método, isso também porque a técnica de eletroforese capilar utilizada para analisar STRs tem baixa sensibilidade [19]. *Single Nucleotide Polymorphisms* (SNPs) posicionados ao longo do genoma são marcadores genéticos adequados para esse tipo de análise, pois têm se mostrado como bons candidatos para a identificação humana [20, 21].

SNP é um tipo de marcador genético que normalmente representa a variação em uma única base da sequência de DNA [22]. Métodos de genotipagem de SNPs são mais rápidos e possuem taxas de erros menores [23, 24]. Outra vantagem de se utilizar SNPs ao invés de STRs é a possibilidade de se analisar DNA degradado e misturas de DNA [16, 23]. Além disso, a taxa de mutação de SNPs, cuja ordem é de 10^{-8} é muito menor do que a de STRs, que é da ordem de 10^{-3} [16, 25]. Porém, ainda existe a necessidade de se analisar muitos *loci* de SNPs porque estes possuem poucos alelos diferentes (entre dois e quatro) [16].

Em um estudo realizado por Guo e colaboradores, SNPs foram identificados usando a plataforma Ion Torrent PGMTM e o kit HID-Ion AmpliSeqTM Identity Panel [26]. O kit foi utilizado em amostras de DNA degradado, misturas de DNA, amostras de pais e filhos e amostras de uma população. Para todos os testes, o kit foi capaz de identificar indivíduos e/ou testar a paternidade [26]. Em um outro estudo de caso, um homem foi encontrado carbonizado e análises utilizando aspectos anatômicos (arcada dentária) e STRs não foram suficientes para fornecer informações acerca da ancestralidade do indivíduo. Utilizaram então o *kit* HID-Ion AmpliSeqTM Ancestry Panel (Thermo Fisher Scientific) para a plataforma Ion Torrent PGMTM, e ele foi capaz de fornecer informações acerca da ancestralidade do indivíduo através de SNPs [27].

Um exame pré-natal não invasivo utilizando *microarrays* já foi desenvolvido por Ryan e colaboradores [8]. Neste exame são analisados aproximadamente 300 mil SNPs em cada uma das três amostras. Levando em conta todas as fontes de erro inerentes à técnica, para cada combinação de mãe e suposto pai, é gerada uma estatística de quão bem o genótipo do suposto pai explica aquele encontrado no componente fetal do fcfDNA plasmático [8]. No entanto, essa técnica ainda possui preços proibitivos para a realidade brasileira e ainda não é realizado no Brasil.

1.3 Micro-haplótipos

O uso de SNPs como marcadores genéticos pode ser limitado pela necessidade de se analisar muitos *loci* (de 40 a 60) para se obter probabilidades como as obtidas em análise de STRs [16]. E as misturas de DNA podem ser difíceis de interpretar por conta da baixa heterozigosidade dos SNPs se comparados com STRs [16]. Nesse contexto, existe a possibilidade de se utilizar micro-haplótipos como marcadores genéticos.

Figura 1.1 - Exemplo da localização do micro-haplótipo no cromossomo. Micro-haplótipos são segmentos cromossomais com um tamanho de até 200 pb, contendo dois a mais SNPs cujas combinações formem pelo menos três haplótipos distintos.

Micro-haplótipos são segmentos cromossomais menores que 200 bp, contendo dois ou mais SNPs que formem pelo menos três haplótipos distintos (Figura 1.1) [21, 23]. A heterozigosidade de um micro-haplótipo está relacionada com a quantidade de haplótipos diferentes que são observados na população. A Tabela 1.1 ilustra um exemplo de micro-haplótipo. Este possui seis haplótipos distintos observados na população, apesar da

combinação desses SNPs poder formar até 192 haplótipos distintos.

Tabela 1.1 - Usando os SNPs ilustrados na Figura 1.1, temos o ID dos SNPs, o alelo referência (REF) e o alternativo (ALT), e os HPs (HP1, HP2, HP3, HP4, HP5 e HP6) representam os haplótipos observados. Temos também SNPs bi-alélicos (rs1, rs8, rs9 e rs10), tri-alélicos (rs3) e tetra-alélicos (rs19).

ID	REF	ALT	HP1	HP2	HP3	HP4	HP5	HP6
rs1	C	A	C	C	A	C	A	C
rs3	G	T,C	G	T	C	G	T	C
rs8	C	G	C	C	G	G	G	G
rs9	C	A	C	A	A	C	A	A
rs10	C	T	C	C	T	C	C	C
rs19	T	A,C,G	T	A	C	T	G	T

Com uma distância de até 10 kb entre SNPs, a taxa de recombinação é da ordem da taxa de mutação dos SNPs [23]. Essa é uma vantagem dos SNPs em relação aos STRs, pois estes possuem taxas de recombinação e mutação maiores que a ordem de 10^{-8} [16, 25]. Dado que a distância entre os SNPs de um micro-haplótipo é menor do que 10 kb, podemos considerar que a taxa de recombinação dos SNPs é menor do que 10^{-8} .

Micro-haplótipos são marcadores genéticos que possuem o potencial de serem utilizados para identificação de indivíduos, inferência da ancestralidade e detecção e quantificação de misturas de material genético [21, 23]. Uma vez que o cfDNA possui um tamanho médio de aproximadamente 145 pb, o seu tamanho é compatível com os micro-haplótipos. Isso significa que a partir de um fragmento desse DNA é possível obter o sequenciamento dos SNPs que compõem o micro-haplótipo [2, 10, 21].

1.4 Sequenciamento de Nova Geração

O método de Sanger é uma tecnologia de sequenciamento baseada na utilização de dideoxi-nucleotídeos e de eletroforese, sendo sua capacidade de sequenciamento de 96 reações em cada processamento [28]. Já o Sequenciamento de Nova Geração (*Next-Generation*

Sequencing, NGS) é uma tecnologia que utiliza métodos diferentes (sequenciamento por síntese ou semicondutor) para fazer o sequenciamento em tempo real, e tem um alto rendimento por sequenciar milhares de moléculas de DNA em paralelo [28, 29]. Existem diversos sequenciadores NGS de bancada, entre os principais estão a plataforma MiSeq da Illumina e o Ion Torrent Personal Genome Machine (PGM) da *Life Technologies* [30, 31].

Tabela 1.2 - Tabela comparativa das duas principais plataformas de sequenciamento NGS de bancada.

Plataforma	Custo do instrumento	Tempo de corrida	Acurácia	Taxa de erro	Tamanho dos reads
Illumina MiSeq	\$128K	27 horas	Maioria > 30Q	0,80%	Até 150 bases
Ion Torrent PGM	\$80K	2 horas	Maioria 20Q	1,71%	200 bases

As duas plataformas já foram comparadas em diversos estudos [30, 31]. O *MiSeq* utiliza nucleotídeos modificados com um fluoróforo diferente para cada tipo de base, e após a adição da base na fita de DNA, os fluoróforos são excitados com laser. Dependendo da base, uma luz de frequência diferente é detectada [30, 32]. Os *reads* gerados nesse sequenciador possuem um tamanho médio de 150 pb, o tempo de processamento é de aproximadamente 27 horas, e a qualidade reportada é maior do que Q30 [30, 31]. Na tecnologia Ion Torrent PGM, conforme as bases são adicionadas à cadeia de DNA, ocorre uma mudança no pH do meio, e essa alteração é utilizada como medida para fazer o *base calling* [33]. Esse sequenciador gera *reads* de tamanho médio de 200 pb, possui um tempo de sequenciamento de aproximadamente duas horas e a maior parte da acurácia reportada é de Q20 [31].

Em um estudo recente realizado por Saba e colaboradores, o sequenciador Ion Torrent PGM foi utilizado no diagnóstico de uma desordem autossômica recessiva em fetos, de maneira não invasiva [9]. Era necessário determinar os haplótipos do pai e da mãe, detectar a presença do fcfDNA no plasma da mãe e diagnosticar o feto. A tecnologia de sequenciamento se mostrou capaz de detectar a presença do DNA fetal e fornecer dados relevantes e informativos a respeito do feto, tornando possível o seu diagnóstico. Além disso, a abordagem de haplotipagem de SNPs foi muito útil para determinar os haplótipos

que o feto herdou dos pais [9].

1.5 Dados do projeto *1000 Genomes*

O projeto *1000 Genomes* [34, 35] foi responsável por elucidar informações de variantes genéticas. Na terceira fase do projeto, foram usados como amostras 2504 indivíduos provenientes de 26 populações (Apêndice A) distribuídas em cinco super-populações : Africana (AFR), Americana (AMR), Leste Asiática (EAS), Européia (EUR) e Sul Asiática (SAS). As análises englobaram SNPs bi-alélicos, SNPs multi-alélicos, *indels* e um conjunto diversificado de variantes estruturais.

No banco *1000 Genomes*, os dados dos genótipos dos indivíduos estão em fase, por isso eles são separados por uma barra vertical (ex.: 1|0). Essa notação indica que sabe-se de qual cromossomo veio o haplótipo (Tabela 1.3). Por exemplo, supondo que um micro-haplótipo seja composto pelos três SNPs (rs4559261, rs12360512, rs4412392). O SNP rs4559261 possui como alelo referência a base T e o alternativo é a base C. O SNP rs12360512 possui como referência o G e como alternativo o T, e por fim, o SNP rs4412392 possui como referência o G e como alternativo o A. Na notação, quando o genótipo contém zero (0), significa que o alelo é a referência, e quando contém um (1), significa que o alelo é o alternativo. Para SNPs tri-alélicos e tetra-alélicos são usados os números dois (2) e três (3) para indicar os outros alelos alternativos. No exemplo da Tabela 1.3, todos os SNPs são bi-alélicos.

Tabela 1.3 - Exemplo de como os dados do *1000 Genomes* são representados e os haplótipos que eles geram. ID representa o identificador do SNP, REF é o alelo referência, ALT é o alelo alternativo, GT é o genótipo encontrado, e Haplótipo 1 e Haplótipo 2 são os dois haplótipos do indivíduo extraídos das informações do genótipo encontrado.

ID	REF	ALT	GT	Haplótipo 1	Haplótipo 2
rs4559261	T	C	0 1	T	C
rs12360512	G	T	1 1	T	T
rs4412392	G	A	1 0	A	G

Portanto, pela Tabela 1.3, sabemos que um dos haplótipo é TTA e o outro haplótipo desse indivíduo é CTG. No entanto, quando o genótipo está separado por uma barra inclinada (1/0), isso significa que não se sabe de qual cromossomo veio o haplótipo e os genótipos não estão em fase (Tabela 1.4). Por exemplo:

Tabela 1.4 - Exemplo da notação utilizada para dados fora de fase e os haplótipos que eles podem gerar. ID representa o identificador do SNP, REF é o alelo referência, ALT é o alelo alternativo, GT é o genótipo encontrado. Os Haplótipos 1, 2, 3 e 4 representam os haplótipos que podem ser formados a partir do genótipo GT.

ID	REF	ALT	GT	Haplótipo 1	Haplótipo 2	Haplótipo 3	Haplótipo 4
rs4559261	T	C	0/1	T	T	C	C
rs12360512	G	T	1/1	T	T	T	T
rs4412392	G	A	1/0	A	G	A	G

Nesse caso, pela Tabela 1.4, existem quatro possíveis haplótipos gerados pela combinação dos alelos dos SNPs. Combinando os haplótipos e levando em conta a informação do genótipo (GT) do indivíduo, temos que o indivíduo pode ter:

- Um haplótipo TTA e outro CTG.
- Um haplótipo TTG e outro CTA.

Objetivo

Apesar dos exames invasivos oferecerem riscos relativamente baixos para a gravidez, estes ainda podem acarretar em um aborto. Dessa forma, há um grande interesse em se utilizar o DNA fetal livre para realizar exames não invasivos. No entanto, os métodos atuais possuem elevado custo dado a enorme quantidade de SNPs a serem analisadas. Para diminuir este custo, propomos utilizar um número de SNPs 1807 vezes menor do que muitos dos métodos atuais. O desafio deste trabalho é, portanto, como obter confiança a partir de um número pequeno de SNPs.

2.1 Objetivo principal

Criar um método de inferência da probabilidade de paternidade em testes pré-natais não invasivos, usando dados de SNPs dentro de micro-haplótipos.

2.2 Objetivos específicos

- Obter os haplótipos do banco de dados *1000 Genomes*.
- Determinar os haplótipos do suposto pai e da mãe.
- Detectar se existe a presença do DNA fetal na amostra de plasma.
- Calcular a probabilidade da paternidade em casos onde não existe a exclusão.

Definição dos micro-haplótipos

Em um teste de paternidade, existe a necessidade de se utilizar informações populacionais para fazer o cálculo da probabilidade do suposto pai ser o pai verdadeiro da criança, considerando casos onde não houve a exclusão da paternidade. Portanto, com o objetivo de se obter as frequências populacionais das regiões que estão sendo analisadas, utilizamos o banco de dados *1000 Genomes*.

3.1 Materiais e Métodos

3.1.1 Micro-haplótipos

Os micro-haplótipos foram previamente escolhidos pelo Dr. Martin Whittle, da empresa Genomic Engenharia Molecular, com base em dois artigos: *Haplotype counting by Next-Generation Sequencing for ultrasensitive human DNA detection* [19] e *Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics* [21].

No artigo *Haplotype counting by Next-Generation Sequencing for ultrasensitive human DNA detection* [19], Debeljak desenvolveu um método bioinformático para detectar regiões polimórficas no genoma, utilizando o banco de dados de quatro populações do *1000 Genomes* (CEU, JPT, CHB e YRI). Os critérios de escolha dessas regiões eram: conter pelo menos nove SNPs dentro de um segmento de 300 pb e o alelo de menor frequência deveria ser $> 9\%$ nas populações analisadas.

Segundo a definição de Kidd, micro-haplótipos são segmentos cromossômicos menores do que 200 pb, contendo dois ou mais SNPs que formam pelo menos três haplótipos distintos. Com base nesta definição e nas características de escolha das regiões encontradas por Debeljak, o Dr. Martin Whittle definiu 20 regiões para serem utilizadas como micro-

haplótipos no estudo. A vantagem das regiões encontradas por Debeljak sobre as regiões que Kidd utiliza para definir seus micro-haplótipos está no número de SNPs utilizados. Enquanto que no trabalho de Debeljak as regiões deveriam conter mais de nove SNPs, no trabalho de Kidd, estas regiões contem pelo menos dois SNPs [19, 21, 23].

Na Tabela 3.1, estão discriminados os 20 micro-haplótipos escolhidos, assim como o cromossomo ao qual pertencem, o intervalo (posição do primeiro SNP e do último SNP) no genoma referência *hg19* e o ID dos SNPs que formam o micro-haplótipo.

Tabela 3.1 - Lista dos 20 micro-haplótipos escolhidos pelo Dr. Martin Whittle para serem analisados. A lista contém uma descrição dos cromossomos ao qual pertencem, a posição inicial e final do micro-haplótipo e os SNPs que estão sendo utilizados dentro do segmento cromossomal.

Nome	Cromossomo	Intervalo (hg19)	Nº de SNPs	ID dos SNPs
M01	4	7447228-7447353	8	rs11721645, rs11729625, rs62277606, rs11729649, rs11729650, rs57770503, rs77178877, rs58285307
M02	4	66995979-66996060	7	rs4342235, rs4395555, rs4365776, rs2882483, rs4339264, rs4621490, rs34805581
M03	5	178259776-178259891	11	rs77700663, rs6894132, rs186140852, rs71611466, rs4700814, rs75686454, rs71611467, rs66505308, rs71611468, rs71611469, rs6879858
M04	6	31319457-31319544	8	rs9266064, rs112974895, rs9266065, rs9405083, rs114945721, rs9266066, rs35370128, rs9266067
M05	8	3478385-3478516	6	rs58829796, rs4875753, rs4875754, rs4875755, rs3102099, rs3110303
M06	8	6160312-6160419	7	rs4559261, rs13260512, rs4412392, rs113523134, rs112950194, rs4615601, rs4282592
M07	9	95691407-95691533	12	rs13296762, rs13300953, rs13296126, rs112492224, rs7027556, rs13294885, rs76237981, rs7027677, rs7028639, rs7027690, rs7028645, rs7027692
M08	10	123095163-123095255	10	rs7899032, rs75102425, rs7913694, rs7913820, rs9421409, rs9421410, rs201989249, rs7913709, rs74158578, rs7913828
M09	10	133376280-133376388	7	rs200254070, rs201475402, rs112519395, rs10830050, rs10830051, rs11018067, rs11018068
M10	11	5078999-5079121	9	rs9804487, rs12806675, rs11035381, rs12790069, rs139300303, rs11035382, rs12789835, rs12789094, rs12789111
M11	13	33553549-33553649	8	rs378609, rs474054, rs9315199, rs60766807, rs2149859, rs570992, rs76400999, rs571057
M12	13	99084196-99084258	6	rs11620100, rs12868939, rs79089579, rs9554471, rs12869473, rs11616733

Continua na próxima página. . .

Tabela 3.1 - Continuação

Nome	Cromossomo	Intervalo (hg19)	Nº de SNPs	ID dos SNPs
M13	14	22736236-22736376	9	rs6572348, rs6572349, rs6572350, rs6572351, rs6572352, rs6572353, rs6572354, rs191987151, rs12880936
M14	15	25047453-25047569	13	rs12914023, rs12914028, rs12914032, rs12914037, rs80258314, rs28864389, rs12914188, rs12915332, rs78817707, rs146871786, rs12900575, rs12915886, rs12914223
M15	16	56576538-56576650	7	rs12444798, rs417053, rs12447596, rs6499839, rs9935553, rs9934924, rs9934927
M16	16	84540654-84540714	7	rs247858, rs12598259, rs28727369, rs12598222, rs12598261, rs171576, rs143449101
M17	17	80804183-80804314	4	rs8069046, rs8064468, rs12945348, rs9912486
M18	18	631314-631435	9	rs11665412, rs11662817, rs11660198, rs11665416, rs28532598, rs13381956, rs11662827, rs11665418, rs10502288
M19	18	76597180-76597277	6	rs4799224, rs4799225, rs4799226, rs4799227, rs4799228, rs4799229
M20	20	1895570-1895673	12	rs66523711, rs66600581, rs73569345, rs112203261, rs111980944, rs6045399, rs111751479, rs113347998, rs113961013, rs114169528, rs114623029, rs76897346

3.1.2 Análise do projeto *1000 Genomes*

Existem diversas formas de se obter os dados do banco *1000 Genomes*, sendo uma destas formas através do *data slicer*, uma ferramenta disponível no site do banco de dados. Através dessa ferramenta, é possível escolher as regiões de interesse, tais como o cromossomo e as regiões dos mesmos. Os dados são fornecidos em arquivos no formato VCF (Figura 3.1).

Os dados fornecidos pelo *1000 Genomes* englobam todas as variantes reportadas que se encontram no intervalo solicitado. Dessa forma, é necessário extrair os dados somente dos SNPs que formam os micro-haplótipos definidos. Usando os dados dos genótipos dos SNPs dos 2504 indivíduos analisados no banco de dados, a frequência relativa de cada haplótipo diferente pode ser calculada. Além dos dados genotípicos, o banco de dados também fornece a anotação das amostras, com informações de população, super-população e gênero.

Assim, é possível saber quais são os haplótipos existentes dentro das populações do banco de dados. Além de obter os haplótipos existentes nas populações do *1000 Genomes*, precisamos das frequências relativas de cada um deles na população, pois esta informação será necessária para se fazer o cálculo da probabilidade de paternidade.

Figura 3.1 - Ferramenta *data slicer* do banco de dados *1000 Genomes*. O *data slicer* é uma ferramenta utilizada para se obter os dados de genótipos do *1000 Genomes*. É possível selecionar apenas as regiões cromossomais onde o micro-haplótipo está localizado e obter os dados no formato VCF, dos 2504 indivíduos que compõem o banco.

3.1.3 Pipeline

Desenvolvemos um *script* para extrair, dos arquivos VCF obtidos do *1000 Genomes*, os haplótipos observados no banco de dados. O *script* buscou nos arquivos VCF os SNPs que formam os micro-haplótipos. Como para cada indivíduo nós temos os dados dos genótipos em fase, sabemos quais são os haplótipos de cada uma das amostras. Utilizando as informações dos cromossomo, da posição dos SNPs, dos IDs e do alelos referência e alternativo, podemos separar os dois haplótipos de cada indivíduo do banco de dados. Dessa forma, podemos calcular a frequência de cada haplótipo levando-se em conta todos

os indivíduos do banco de dados, ou uma dada super-população ou população.

Na Tabela 3.2, podemos observar os genótipos de três indivíduos do *1000 Genomes*, sendo zero (0) o alelo referência e um (1) o alelo alternativo. Temos então que o indivíduo HG00096 tem um haplótipo CCGT e outro TCGT, o indivíduo HG00097 tem um haplótipo CCAC e CCGT e o indivíduo HG00099 tem um haplótipo TCAC e CCGT.

Tabela 3.2 - Disposição dos dados do vcf no micro-haplótipo M17. Os dados estão separados por cromossomo, posição inicial, ID da variante, alelo referência, alelo alternativo, e indivíduos.

CHROM	POS	ID	REF	ALT	HG00096	HG00097	HG00099
17	80804183	rs8069046	T	C	1 0	1 1	0 1
17	80804210	rs8064468	C	T	0 0	0 0	0 0
17	80804285	rs12945348	G	A	0 0	1 0	1 0
17	80804314	rs9912486	C	T	1 1	0 1	0 1

3.2 Resultados e Discussão

O resultado do *script* desenvolvido segue na forma de tabelas. Abaixo, na Tabela 3.3, temos a lista de haplótipos do micro-haplótipo M01, assim como a frequência de cada um dos haplótipos observados, levando-se em conta todos os indivíduos do banco de dados, e os indivíduos das super-populações AFR, AMR, EAS, EUR e SAS. No Apêndice B estão as tabelas com a lista dos outros 19 micro-haplótipos analisados nesse trabalho.

Tabela 3.3 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M01 e a frequência deles levando em conta todas as populações (TODOS) e as super-populações (AFR, AMR, EAS, EUR e SAS).

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M01H01	AAAAGGCA	1	0	0	0	1	0
M01H02	AAAGGACG	1	0	0	0	0	1
M01H03	AAAGGGCA	1344	73	277	270	368	356

Continua na próxima página...

Tabela 3.3 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M01H04	AAAGGGCG	23	0	1	8	10	4
M01H05	AAAGGGTA	118	39	20	0	34	25
M01H06	AAGAAACA	9	1	1	3	4	0
M01H07	AAGAAACG	1700	322	183	557	286	352
M01H08	AAGAAGCA	2	0	0	1	0	1
M01H09	AAGAGACA	1	1	0	0	0	0
M01H10	AAGAGACG	1	0	0	0	0	1
M01H11	AAGGGACG	8	8	0	0	0	0
M01H12	AAGGGGCG	48	3	15	0	23	7
M01H13	AGAGGGCA	1	0	0	0	1	0
M01H14	AGGAGACA	114	106	7	0	1	0
M01H15	AGGGGACG	1	0	1	0	0	0
M01H16	AGGGGGCG	1	1	0	0	0	0
M01H17	CAAGGGCA	1	0	1	0	0	0
M01H18	CAGAAACG	3	0	0	2	1	0
M01H19	CAGGGACG	1	0	0	0	1	0
M01H20	CGAGGACG	1	1	0	0	0	0
M01H21	CGAGGGCA	14	12	1	0	1	0
M01H22	CGAGGGTA	1	1	0	0	0	0
M01H23	CGGAAACG	1	0	0	1	0	0
M01H24	CGGAAGCA	9	9	0	0	0	0
M01H25	CGGGGACA	99	96	3	0	0	0
M01H26	CGGGGACG	620	206	81	143	132	58
M01H27	CGGGGGCA	317	297	15	0	5	0
M01H28	CGGGGGCG	568	146	88	23	138	173

Podemos observar na Tabela 3.3 que existem alguns haplótipos que estão presentes em apenas uma das super-populações (M01H01, M01H02 e M01H09). Existem haplótipos que são observados em 80% das super-populações (M01H04, M01H05 e M01H06). E por fim, haplótipos que são encontrados em todas as super-populações (M01H03, M01H07 e M01H26). O mesmo padrão encontrado nesse micro-haplótipo pode ser observado nos outros micro-haplótipos (ver Apêndice B), e isso pode ser um indicativo de que os SNPs e haplótipos possuem informação de ancestralidade.

Com o objetivo de verificar se existe informação de ancestralidade nos SNPs analisados, utilizamos os dados do *1000 Genomes* para realizar uma Análise de Componente Principal (PCA, *Principal Component Analysis*) com as informações dos 2504 indivíduos do banco.

Utilizamos as informações de 164 SNPs bi-alélicos que compõem os micro-haplótipos analisados (ver Tabela 3.1). Nessa análise, esperamos que, se houver informação de ancestralidade, os indivíduos das super-populações estarão agrupados em conjuntos e estes estarão separados e distantes uns dos outros. O resultado se encontra na Figura 3.2.

Figura 3.2 - Gráfico do PC1 versus o PC2 dos SNPs analisados. O PCA foi realizado com os dados de 164 SNPs dos 2504 indivíduos do *1000 Genomes*. Os indivíduos foram separados pelas super-populações: AFR, AMR, EAS, EUR e SAS. E observamos que existe uma pequena separação entre as super-populações.

Apesar de não ser observada uma separação total, podemos ver na Figura 3.2 que existe uma pequena separação entre os indivíduos das cinco super-populações (AFR, AMR, EAS, EUR e SAS). Isso acontece porque apesar de estarmos analisando 164 SNPs, eles estão próximos uns dos outros e agrupados dentro de 20 micro-haplótipos. Dessa forma, embora a separação não seja completa, podemos concluir que existe informação de ancestralidade nesses SNPs.

Análise dos dados de sequenciamento

Em um teste de paternidade pré-natal não invasivo, é necessário coletar amostras sanguíneas de dois indivíduos: o suposto pai e a mãe grávida. A amostra sanguínea da mãe é então processada e separada em duas amostras: a fração celular (DNA materno apenas) e o plasma (mistura de DNA livre materno e fetal). Ao fazer o sequenciamento dessas três amostras de DNA (suposto pai, mãe e plasma), podemos determinar os haplótipos do suposto pai e da mãe. A amostra de sequenciamento do plasma é utilizada para detectar a presença de DNA fetal livre e determinar o haplótipo que o feto herdou do pai biológico.

4.1 Materiais e Métodos

4.1.1 Genomic Engenharia Molecular

A empresa Genomic Engenharia Molecular, representada pelo Dr. Martin Whittle, foi a responsável por coletar as amostras sanguíneas, processá-las e realizar o sequenciamento e pré-processamento dos dados. O sequenciamento de painel das amostras foi feito utilizando o sequenciador Ion Torrent PGM e o pré-processamento foi realizado pelo *Torrent Suite Software*, fornecendo os dados no formato BAM. Após estas etapas, os arquivos BAM dos trios de amostras (suposto pai, mãe e plasma) foram disponibilizadas no *site* da empresa (<http://genomic.com.br/banco-de-dados/>). Utilizando esse banco de dados público de amostras de trios, desenvolvemos um *script* e fizemos as nossas análises.

O formato de arquivo BAM é a versão binária do formato SAM (*Sequence Alignment/Map*). As informações de sequenciamento desse arquivo estão dispostos em forma de colunas e linhas. Cada linha corresponde a um *read* do sequenciamento, e cada coluna dessa linha contém uma informação a respeito do *read*. Na Tabela 4.1, temos a coluna do

arquivo BAM/SAM, seu nome e a descrição da informação armazenada [36, 37].

Tabela 4.1 - Cada coluna dos arquivos BAM e SAM armazena uma informações diferente a respeito do *read*.

Coluna	Nome	Descrição
1	QNAME	Nome do <i>read</i> ou par de <i>reads</i>
2	FLAG	<i>Bits</i> indicando diversas informações sobre o alinhamento
3	RNAME	Nome da sequência na referência
4	POS	Posição mais à esquerda do <i>read</i> que se alinha na referência
5	MAPQ	Qualidade do Mapeamento (Escala Phred)
6	CIGAR	CIGAR <i>string</i>
7	MRNM	Referência do próximo <i>read</i> no par/segmento
8	MPOS	Posição do próximo <i>read</i> no par/segmento
9	ISIZE	Tamanho do template observado
10	SEQ	Sequência do <i>read</i> na mesma fita da referência
11	QUAL	Qualidade do <i>read</i> (ASCII-33 = Qualidade de base Phred)
12	FIELDS	Campos opcionais

4.1.2 Bioinformática

Os arquivos BAM do suposto pai e da mãe devem passar por uma etapa de qualidade para depois serem genotipados. O arquivo do plasma também é analisado qualitativamente. Por fim, as informações dos genótipos do suposto pai e da mãe são analisados na estatística juntamente com o resultado da qualidade do plasma. Para incluir as informações populacionais na estatística, utilizamos o banco de dados de haplótipos criado a partir do *1000 Genomes* (ver no Capítulo 3). Na Figura 4.1 apresentamos um *workflow* das etapas desenvolvidas no projeto.

Os dados de sequenciamento estão no formato BAM, e para obter os haplótipos dos indivíduos, foi utilizado um pipeline baseado em um dos componentes do pacote do SAMtools: samtools. O SAMtools [37] é uma ferramenta utilizada para manipular arquivos no formato BAM, permitindo fazer análises de dados de sequenciamento genômico. Esta ferramenta é capaz de converter formatos de alinhamentos, ordenar e unir alinhamentos,

remover duplicatas de PCR, detectar SNPs e indels, etc [36, 37]. Usando a linguagem de programação *Perl*, desenvolvemos um conjunto de *scripts* que utiliza a ferramenta SAMtools [37] para manipular os arquivos BAM e extrair deles as informações de cada *read*. Após a extração dessas informações, desenvolvemos *scripts* em *Perl* para manusear os dados e obter as informações de haplótipos e frequências.

Figura 4.1 - Workflow do processamento e análise dos dados. Os arquivos BAM da mãe e do suposto pai são filtrados na etapa da *Qualidade*, e o resultado é analisado na etapa da *Genotipagem*, onde os dois indivíduos têm o seu genótipo determinado. O arquivo BAM do plasma é filtrado na etapa da *Qualidade* e o seu resultado é analisado na etapa *Probabilidade de Paternidade* (ver no Capítulo 5). Na etapa *Probabilidade de Paternidade*, são então analisados os genótipos da mãe e do suposto pai, o resultado da qualidade do arquivo BAM do plasma e as informações populacionais obtidas no banco de dados de haplótipos (ver no Capítulo 3).

Qualidade

Para analisar a qualidade dos dados, cada *read* do arquivo BAM analisado deve passar por uma série de etapas (Figura 4.2). Primeiramente, verificamos se o *read* está alinhado no intervalo correspondente ao micro-haplótipo ($M[i]$) analisado. Posteriormente, verificamos se o *read* está alinhado em apenas uma região ou duas. Queremos *reads* que estejam alinhados em apenas um região. Após esta etapa, analisamos a qualidade do mapeamento do *read*, se a qualidade for maior do que 20 (escala Phred), o *read* segue para as análises posteriores, caso contrário, ele é descartado.

Após analisar o mapeamento, analisamos o CIGAR *string*. Esse parâmetro contém diversas informações a respeito do alinhamento e mapeamento do *read* no genoma referência, tais como inserções, deleções, *matches* e *mismatches*. Desejamos *reads* cujo CIGAR *string* contenha apenas *match* e *mismatch*, que são representados pela letra M.

Nessa etapa, para determinar os haplótipos dos indivíduos analisados, precisamos obter as informações das bases que correspondem aos SNPs analisados em um micro-haplótipo. Como sabemos a posição dos SNPs, extraímos das informações dos *reads* as bases corres-

pondentes à essas posições, ver na Figura 4.3. Além disso, nas informações de sequenciamento, também existe a qualidade do sequenciamento da base. Dessa forma, temos o haplótipo encontrado no *read* e a qualidade de cada uma das bases que constitui o haplótipo.

Figura 4.2 - Workflow da etapa da Qualidade. Como mostrado na Figura 4.1, os arquivos BAM da mãe e do suposto pai passam pela etapa da Qualidade e a lista de possíveis haplótipos é utilizada para fazer a Genotipagem. Após esta etapa, os genótipos do suposto pai e da mãe são utilizados na Probabilidade de Paternidade. O arquivo BAM do plasma passa pela etapa da Qualidade e a lista de possíveis haplótipos é utilizada na etapa Probabilidade de Paternidade.

Figura 4.3 - Extração das informações dos SNPs dos reads. Como as posições dos SNPs no genoma são conhecidas, podemos extrair as informações das bases e formar o haplótipo.

Se todas as bases, que formam o haplótipo em um dado *read*, possuírem uma qualidade acima de um escore, o haplótipo é então adicionado a uma lista. Se alguma base do haplótipo possuir uma qualidade abaixo do escore, essa base é então substituída por um traço (-), e esse traço indica que a base é desconhecida, conforme ilustrado na Figura 4.4. Para as nossas análises, o valor do *threshold* utilizado foi de 20 (escala Phred).

Figura 4.4 - Qualidade das bases e determinação de haplótipos. A determinação dos haplótipos dos *reads* depende do escore da qualidade utilizado. Se a qualidade da base for menor do que o valor de corte, a base é substituída por um traço (-), e a sua informação se torna desconhecida.

Figura 4.5 - Pareamento dos *reads* em haplótipos. Os *reads* que possuem mais de 30% de traços na sua composição são excluídos, e os *reads* que possuem qualidade são utilizados. Os *reads* que passaram na etapa anterior são organizados em uma lista, e dessa lista são selecionados os haplótipos únicos. Todos os *reads* são então pareados nos haplótipos únicos, e caso um *read* pareie em mais de um haplótipo, ele é descartado para não adicionar informação redundante. Após o pareamento, são somadas a quantidade de *reads* que parearam em cada haplótipo único.

Os *reads* que possuem haplótipos com traços devem passar por um outro controle de qualidade. Os haplótipos precisam conter mais de 70% de bases conhecidas, ou seja, podem conter no máximo 30% de traços na sua composição, ver Figura 4.5. Caso o *read* passe

por esse controle de qualidade, esse *read* é pareado aos haplótipos que tenham todas as suas bases conhecidas.

Na lista de haplótipos, caso um *read* esteja pareado em mais de um haplótipo, ele é então descartado, pois a adição dele na contagem pode produzir informação redundante, como ilustrado na Figura 4.5. Por fim, obtemos uma lista de possíveis haplótipos para cada um dos 20 micro-haplótipos da amostra analisada. Essa lista será utilizada para determinar os genótipos do suposto pai e da mãe (Genotipagem) ou para detectar os possíveis haplótipos que o feto herda somente do suposto pai.

Genotipagem

Precisamos saber qual o genótipo do suposto pai e da mãe para prosseguir com as análises. Então, para cada um dos 20 micro-haplótipos, utilizamos a lista de possíveis haplótipos obtida na etapa da qualidade. Com base em uma revisão de literatura intitulada *Genotype and SNP calling from next-generation sequencing data* [38], para um SNP ter qualidade suficiente para ser analisado, a cobertura dele tem que ser maior do que 20X. Isso significa que é necessário que hajam 20 *reads* do sequenciamento cobrindo a posição do SNP. Além disso, o artigo também sugere que um SNP pode ser considerado heterozigoto se houver um desbalanço de 20% e 80%. Isso significa que, caso um SNP tenha uma contagem de 20 *reads* para A e uma contagem de 80 *reads* para C, ele é considerado heterozigoto nessa posição, e um dos alelos é A e o outro é C. Dessa forma, utilizando o mesmo raciocínio, escrevemos um *script* onde, dado uma lista de haplótipos, queremos determinar qual o genótipo do indivíduo analisado. Dado que a diferença entre dois haplótipos diferentes da lista seria de no mínimo um SNP.

Nesta etapa, para cada micro-haplótipo, utilizamos a lista de haplótipos para determinar o genótipo deste *locus*. Determinamos que, se houver um haplótipo com uma frequência relativa maior do que 80%, o genótipo será homozigoto para esse haplótipo. Caso não haja um haplótipo com frequência relativa maior do que 80%, analisamos se existem dois haplótipos com frequência relativa entre 20% e 80%. Em caso positivo, o indivíduo será heterozigoto para os dois haplótipos. Em caso negativo, concluímos que a qualidade não é boa para determinar o genótipo. Conforme pode ser visto na Tabela 4.2.

Tabela 4.2 - Regras para determinar o genótipo da mãe ou do suposto pai com base na lista de haplótipos.

Regra	Significado
Um haplótipo > 80%	Homozigoto
Dois haplótipos entre 20% e 80%	Heterozigoto

4.2 Resultados e Discussão

4.2.1 Simulação dos dados do *1000 Genomes*

Figura 4.6 - Simulação utilizando os dados do *1000 Genomes*. Seleccionamos um indivíduo masculino e um feminino, extraímos os seus genótipos, e criamos um genótipo para o filho. Em amarelo, simulamos a paternidade verdadeira, onde testamos a paternidade utilizando o pai verdadeiro. Em azul, simulamos a paternidade falsa, onde testamos a paternidade com um terceiro indivíduo que sabemos que não é o pai verdadeiro.

Com o objetivo de analisar o impacto do número de micro-haplótipos no valor da probabilidade de paternidade, realizamos algumas simulações utilizando os dados do *1000 Genomes*. Primeiramente, buscamos analisar a relação entre o número de micro-haplótipos e a precisão do cálculo da probabilidade de paternidade. Para isso, seleccionamos do banco de dados um indivíduo masculino e um feminino para serem os pais biológicos. Extraímos os seus genótipos para cada um dos 20 micro-haplótipos e criamos um genótipo para o filho, que recebeu um haplótipo do pai e um da mãe (ver a Figura 4.6).

Após essa etapa, utilizamos as informações populacionais dos haplótipos, e obtivemos a probabilidade de paternidade utilizando um micro-haplótipo, dois micro-haplótipos, e assim por diante até utilizar os 20. A simulação foi repetida 26.000 vezes usando todo o banco de dados. Fizemos a mesma simulação com uma diferença, ao invés de utilizar todo o banco de dados, utilizamos uma das 26 populações e fizemos 1.000 repetições, e repetindo o mesmo procedimento para as outras 25 populações, obtivemos 26.000 repetições. Ambos os resultados estão apresentados em amarelo no gráfico da Figura 4.7 e estão separados em “Por população”, onde as simulações foram feitas utilizando uma população específica, e “Todo o banco”, onde todos os dados do *1000 Genomes* foram utilizados.

Figura 4.7 - Boxplot da probabilidade de paternidade versus o número de micro-haplótipos utilizados para o cálculo da probabilidade. Para cada situação, 26.000 repetições da simulação foram realizadas. *Falso* indica a situação onde o suposto pai testado não é o pai verdadeiro, e *Verdadeiro* indica a situação onde o suposto pai testado é o pai verdadeiro. “Por população” são as simulações feitas utilizando uma população específica de onde os indivíduos foram amostrados, e “Todo o banco” são as simulações feitas utilizando todos os dados do *1000 Genomes* para amostrar os indivíduos. Para o caso *Verdadeiro*, esperamos valores de probabilidade próximos de um, e no caso *Falso*, esperamos valores próximos de zero.

Com o objetivo de descobrir a quantidade de micro-haplótipos necessários para indicar a exclusão de um suposto pai não verdadeiro, fizemos uma simulação de paternidade falsa. Da mesma forma como descrito na simulação anterior, criamos um filho artificial (ver a Figura 4.6). No entanto, selecionamos um outro indivíduo masculino para testar a pater-

nidade, e sabemos que ele não deverá ser incluso como pai biológico. Essa simulação foi repetida 26.000 vezes usando todo o banco de dados. Fizemos as simulações de paternidade falsa utilizando uma das populações do *1000 Genomes* por vez, e repetindo 1.000 vezes para cada população, resultando 26.000 simulações. O resultado de ambas as situações está apresentado em azul no gráfico da Figura 4.7, onde “Por população” indica as simulações realizadas com uma população específica e “Todo o banco” indica as simulações realizadas com o todos os dados do *1000 Genomes*.

Podemos observar a partir dos resultados das simulações de paternidade verdadeira, apresentados na Figura 4.7 (em amarelo), que quanto mais micro-haplótipos analisados, maior o valor da Probabilidade de Paternidade. Observando os resultados das simulações de paternidade falsa na Figura 4.7 (em azul), observamos que poucos micro-haplótipos analisados podem resultar em uma falsa inclusão de paternidade. Fizemos um gráfico da quantidade de falsos positivos *versus* o número de micro-haplótipos utilizados na análise (ver Figura 4.8). Consideramos inclusão de paternidade valores de probabilidade maiores ou iguais a 99%.

Figura 4.8 - Dotplot do número de falso positivo em 26.000 simulações *versus* o número de micro-haplótipos utilizados na análise. “Por população” são as simulações feitas utilizando uma população específica de onde os indivíduos foram amostrados, e “Todo o banco” são as simulações feitas utilizando todos os dados do *1000 Genomes* para amostrar os indivíduos. Falsos positivos são os casos onde o valor da probabilidade de paternidade era maior ou igual a 99%. Observamos que conforme aumentamos a quantidade de micro-haplótipos utilizados, a quantidade de falsos positivos diminui.

Concluimos com as simulações anteriores que ao utilizarmos uma quantidade maior de micro-haplótipos, as chances de ocorrerem falsos positivos diminui e a probabilidade de

paternidade assume valores maiores, aumentando a precisão do cálculo.

4.2.2 Análise dos dados da Genomic Engenharia Molecular

No banco de dados de amostras da Genomic Engenharia Molecular, obtivemos 60 trios de amostras do tipo suposto pai, mãe e plasma. O *workflow* foi realizado para todas as amostras. Para ilustrar alguns dos resultados da etapa da qualidade e da genotipagem, utilizamos as amostras T48S01, que corresponde à amostra do suposto pai do trio 48 e a amostra T58S02, que corresponde à amostra da mãe do trio 58.

O primeiro caso, é o micro-haplótipo M02 da amostra T58S02. A lista dos haplótipos encontrados nos dados do sequenciamento, após a etapa da qualidade, estão apresentados na Tabela 4.3.

Tabela 4.3 - Lista dos possíveis haplótipos encontrados no micro-haplótipo M02 da amostra T58S02, essa é a amostra da mãe, do trio 58. Para cada haplótipo encontrado, temos a quantidade de *reads* que parearam em cada haplótipo e a frequência relativa de cada um.

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
GATAAAC	310	49,60%	GGCGCGT	1	0.16%
AGCGCGT	300	48,00%	AGCGCGC	1	0.16%
AGCGCAC	5	0.80%	AGCGCTT	1	0.16%
AGCAAAC	4	0.64%	AATAAAC	1	0.16%
GATGCGT	2	0.32%			

O segundo caso é o micro-haplótipo M03 da amostra T58S02. A lista dos haplótipos encontrados nos dados do sequenciamento, após a etapa da qualidade, está apresentada na Tabela 4.4.

Tabela 4.4 - Lista dos possíveis haplótipos encontrados no micro-haplótipo M03 da amostra T58S02, essa é a amostra da mãe, do trio 58. Para cada haplótipo encontrado, temos a quantidade de *reads* que parearam em cada haplótipo e a frequência relativa de cada um.

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
AGGTCAGTAAA	499	68,55%	ACGTTAGTAAA	2	0,27%
ACGTTAGTAAG	211	28,98%	AGGTCAGTAGA	2	0,27%
AGGTTAGTAAG	5	0,69%	ACGTTAGTGAG	1	0,14%
AGGTCAGTAAG	3	0,41%	ACGCTAGTAAG	1	0,14%
GGGTCAGTAAA	3	0,41%	AGGCCAGTAAA	1	0,14%

O terceiro caso, é o micro-haplótipo M18 da amostra T58S02. A lista dos haplótipos encontrados nos dados do sequenciamento, após a etapa da qualidade, está apresentada na Tabela 4.5.

Tabela 4.5 - Lista dos possíveis haplótipos encontrados no micro-haplótipo M18 da amostra T58S02, essa é a amostra da mãe, do trio 58. Para cada haplótipo encontrado, temos a quantidade de *reads* que parearam em cada haplótipo e a frequência relativa de cada um.

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
CGACCCGCG	1084	99,26%	CGACCGGCG	1	0,09%
CGGCCCCGCG	2	0,19%	CGACCTGCG	1	0,09%
CGACTCGCG	2	0,19%	CGACACGCG	1	0,09%
CGACGCGCG	1	0,09%			

O quarto caso, é o micro-haplótipo M15 da amostra T58S02. A lista dos haplótipos encontrados nos dados do sequenciamento, após a etapa da qualidade, está apresentada na Tabela 4.6.

Tabela 4.6 - Lista dos possíveis haplótipos encontrados no micro-haplótipo M15 da amostra T58S02, essa é a amostra da mãe, do trio 58. Para cada haplótipo encontrado, temos a quantidade de *reads* que parearam em cada haplótipo e a frequência relativa de cada um.

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
GCTCGCA	199	89.64%	GCTCGCG	23	10.36%

O quinto caso, é o micro-haplótipo M15 da amostra T48S01. A lista dos haplótipos encontrados nos dados do sequenciamento, após a etapa da qualidade, está apresentada na Tabela 4.7.

Tabela 4.7 - Lista dos possíveis haplótipos encontrados no micro-haplótipo M15 da amostra T48S01, essa é a amostra do suposto pai, do trio 48. Para cada haplótipo encontrado, temos a quantidade de *reads* que parearam em cada haplótipo e a frequência relativa de cada um.

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
GCTCGCA	40	45,45%	ACTCGCG	9	10,23%
ACTCGCA	36	40,91%	GCTCGCG	3	3,41%

E por fim, o sexto caso, é o micro-haplótipo M13 da amostra T48S01. A lista dos haplótipos encontrados nos dados do sequenciamento, após a etapa da qualidade, está apresentada na Tabela 4.8.

Tabela 4.8 - Lista dos possíveis haplótipos encontrados no micro-haplótipo M13 da amostra T48S01, essa é a amostra do suposto pai, do trio 48. Para cada haplótipo encontrado, temos a quantidade de *reads* que parearam em cada haplótipo e a frequência relativa de cada um.

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
GCTCCACGC	24	30,00%	GTGTTGTGT	4	5,00%
GTGTTGTGC	24	30,00%	GCTCCACGT	3	3,75%

Continua na próxima página...

Tabela 4.8 - Continuação

Haplótipo	Nº de reads	Frequência relativa	Haplótipo	Nº de reads	Frequência relativa
ACTCCACGC	16	20,00%	GTGTTGTAC	1	1,25%
ATGTTGTGC	8	10,00%			

Sabemos que os indivíduos podem ser homozigotos ou heterozigotos, portanto, esperamos encontrar nas listas um ou dois haplótipos distintos. No entanto, conforme pode ser observado nos casos apresentados anteriormente, existem mais haplótipos do que o esperado nessas listas, e provavelmente eles são originados de erros de sequenciamento. Por isso, fizemos um histograma da frequência relativa dos haplótipos encontrados nos dados de sequenciamento das 120 amostras de mãe e suposto pai (Figura 4.9).

Figura 4.9 - Histograma da frequência relativa dos haplótipos encontrados nos sequenciamentos. Utilizando os dados de suposto pai e mãe dos 60 trios de amostra do banco de dados da Genomic, fizemos um histograma da frequência relativa de todos os haplótipos obtidos do sequenciamento das amostras. Observamos que as maiores concentrações estão próximas de 0%, 50% e 100%.

Observamos no gráfico que existe uma grande quantidade de haplótipos em torno de 50% e em torno de 100%. Isso indica que são os haplótipos de *locus* homozigoto e heterozigoto. Mas podemos observar que existe uma variação nessas frequências, por isso precisamos levar em conta o desbalanço que ocorre inerentemente ao método de sequenciamento.

Pelas regras de genotipagem apresentadas da Tabela 4.2, esperamos haplótipos com frequências relativas maiores do que 20%. Observamos que existe uma grande quantidade de haplótipos com frequências relativas menores do que 20% no gráfico da Figura 4.9. Esses provavelmente são oriundos de erros de sequenciamento, como pudemos observar nos casos apresentados.

No primeiro caso e no segundo caso (apresentados nas Tabelas 4.3 e 4.4) observamos que existem dois haplótipos com frequências relativas mais altas. No entanto, o desbalanço do primeiro caso é de 49,60% e 48,00%, e no segundo caso é de 68,55% e 28,98%. Em ambos os casos, os erros de sequenciamento não ultrapassam 10%. Apesar de existir o desbalanço, podemos concluir que ambos são heterozigotos para os haplótipos de frequências relativas mais altas.

No terceiro e quarto caso (apresentados nas Tabelas 4.5 e 4.6) observamos que existe um haplótipo com frequência relativa mais alta. No entanto, no quarto caso (Tabela 4.6) existe um haplótipo com frequência relativa mais alta do que o esperado para erros de sequenciamento (10,36%). Porém, o desbalanço encontrado não cumpre a regra para ser heterozigoto (dois haplótipos entre 20% e 80%), por isso, determinamos que esse tipo de caso será considerado homozigoto se o haplótipo de maior frequência relativa estiver acima de 80%.

No quinto e sexto caso (apresentados nas Tabelas 4.7 e 4.8) observamos que existem dois haplótipos com frequências relativas mais altas (45,45%, 20,91% e 30%). No entanto, também encontramos haplótipos com frequências relativas acima do esperado para erros de sequenciamento (20,00%, 10,23% e 10%). No quinto caso (Tabela 4.7, observamos que existe o erro de sequenciamento, entretanto, os haplótipos de maior frequência relativa estão próximos de 40%. E no sexto caso, os erros são de 20,00% e 10,00% e os haplótipos de maior frequência relativa são de 30,00%, nessa situação, não existe confiabilidade nesse tipo de dado.

Pelas simulações realizadas utilizando os dados do *1000 Genomes*, sabemos que precisamos de mais do que 11 micro-haplótipos com qualidade para diminuirmos as chances de ocorrerem falsos positivos. Portanto, considerando os casos apresentados (casos de um a seis), e aceitando que os erros de sequenciamento ocorrem com frequência relativa abaixo de 10% reformulamos as regras de genotipagem, apresentada na Tabela 4.9.

Tabela 4.9 - Novas regras para determinar o genótipo da mãe ou do suposto pai com base na lista de haplótipos.

Condição 1	Condição 2	Significado
Um haplótipo > 10%	Um haplótipo > 80%	Homozigoto
Dois haplótipos > 10%	Um haplótipo > 80%	Homozigoto
Dois haplótipos > 10%	Dois haplótipos entre 20% e 80%	Heterozigoto
Três haplótipos > 10%	Dois haplótipos > 35%	Heterozigoto

De acordo com o artigo *Non-invasive prenatal diagnosis of beta-thalassemia by semi-conductor sequencing: a feasibility study in the sardinian population* [9], a amostra do plasma precisa ter uma cobertura de pelo menos 1000X no *locus* para que a análise seja confiável. Por isso, realizando a etapa da qualidade em todas as amostras, e exigindo que a amostra do plasma possua uma cobertura de 1000X em cada micro-haplótipo, criamos a Tabela 4.10. Nessa tabela, temos a quantidade de micro-haplótipos que foram genotipados no suposto pai e na mãe (SP e M), e a quantidade de micro-haplótipos que possuem uma cobertura maior do que 1000X no plasma (PL). Além disso, apresentamos a quantidade de micro-haplótipos que estão sendo cobertos em comum às três amostras ($SP \cap M \cap PL$), dessa forma, temos a quantidade de micro-haplótipos que o trio possui com qualidade para ser analisado na etapa seguinte.

Tabela 4.10 - Lista da quantidade de micro-haplótipos cobertos em cada trio e por amostra: suposto pai (SP), mãe (M) e plasma (PL). Apresentamos também a quantidade de micro-haplótipos que estão sendo cobertos em comum a todas as três amostras analisadas no trio ($SP \cap M \cap PL$).

Trio	SP	M	PL	$SP \cap M \cap PL$	Trio	SP	M	PL	$SP \cap M \cap PL$	Trio	SP	M	PL	$SP \cap M \cap PL$
T01	1	0	0	0	T02	0	1	0	0	T03	13	9	1	1
T04	14	10	2	2	T05	9	11	0	0	T06	6	10	0	0
T07	4	5	1	1	T08	5	1	1	1	T09	19	18	19	18
T10	17	19	14	13	T11	18	17	16	16	T12	16	18	17	15
T13	17	16	6	6	T14	16	15	14	13	T15	14	15	13	10
T16	16	16	13	13	T17	16	17	12	11	T18	17	17	12	12
T19	2	2	2	2	T20	2	1	1	1	T21	2	2	1	1

Continua na próxima página...

Tabela 4.10 - Continuação

Trio	SP	M	P	SP $\cap$ M $\cap$ P	Trio	SP	M	P	SP $\cap$ M $\cap$ P	Trio	SP	M	P	SP $\cap$ M $\cap$ P
T22	2	1	1	0	T23	0	0	2	0	T24	0	1	1	0
T25	0	1	0	0	T26	0	1	0	0	T27	0	0	0	0
T28	1	1	2	1	T29	0	1	0	0	T30	0	0	0	0
T31	18	17	18	15	T32	19	18	18	17	T33	16	18	18	15
T34	18	19	17	16	T35	19	18	14	14	T36	8	9	15	0
T37	19	19	18	18	T38	16	13	12	9	T39	16	18	18	16
T40	19	16	18	16	T41	17	16	17	15	T42	18	17	16	16
T43	18	19	19	18	T44	19	18	18	17	T45	15	13	16	13
T46	14	16	18	14	T47	15	13	17	13	T48	18	14	14	12
T49	17	15	17	14	T50	17	15	17	14	T51	18	15	19	15
T52	16	17	17	15	T53	19	17	18	17	T54	16	17	18	16
T55	17	18	18	16	T56	17	16	18	15	T57	17	18	19	16
T58	17	18	18	17	T59	18	18	16	15	T60	17	18	17	15

Para realizarmos o cálculo da probabilidade de paternidade (Capítulo 5), precisamos saber os genótipos do suposto pai e da mãe, e observar se nos dados do plasma existem indícios de que o suposto pai seja o pai verdadeiro. Como visto anteriormente, a partir de 15 micro-haplótipos, as chances de ocorrerem falsas inclusões de paternidade diminuem. Portanto, quanto maior a quantidade de micro-haplótipos cobertos com qualidade nas amostras do suposto pai, mãe e plasma, maiores as chances de determinarmos corretamente a inclusão ou exclusão da paternidade.

Cálculo da probabilidade de paternidade

A probabilidade de paternidade é calculada utilizando um conjunto de *loci* independentes entre si. Juntos, esses *loci* fornecem o poder estatístico necessário para o cálculo da probabilidade de paternidade. Para realizar tal cálculo, precisamos de dados populacionais, e os genótipos da mãe, do feto e do suposto pai.

Nas etapas anteriores, obtivemos os dados populacionais, a partir do *1000 Genomes* e os genótipos da mãe e do suposto pai. Nesta etapa, utilizamos os dados do plasma para obter informação a respeito do feto. Para medir estatisticamente a possibilidade da informação observada no plasma ser de fato fetal e não de erros de sequenciamento, utilizamos as evidências de paternidade.

Em cada *locus*, de acordo com o que for observado nas evidências de paternidade, calculamos o índice de paternidade. Por fim, utilizando os índices de paternidade calculados para cada micro-haplótipo, calculamos a probabilidade de paternidade.

5.1 Materiais e Métodos

5.1.1 Evidências de paternidade

No capítulo 4, mostramos que, para uma dada amostra, após a etapa da checagem da qualidade, obtemos uma lista de possíveis haplótipos para cada *locus* analisado (ver Figura 4.2). Nos dados de sequenciamento da mãe e do suposto pai, esperamos encontrar no máximo dois haplótipos diferentes para cada *locus*. No entanto, devido a erros inerentes à técnica de sequenciamento, a lista de possíveis haplótipos é sempre maior do que o esperado. Para isso, utilizamos um critério de corte para determinar o genótipo de cada micro-haplótipo do indivíduo analisado, conforme visto na Tabela 4.9.

Na herança genética, metade da informação é herdada do pai e a outra é herdada da mãe. Dessa forma, nos dados de sequenciamento do plasma, temos a informação genética do feto misturada com a da mãe. Nesses mesmos dados, serão encontrados um ou dois haplótipos com frequência relativa mais alta (a mãe pode ser homozigota ou heterozigota) e outros haplótipos com frequências relativas menores. O haplótipo que o feto herda da mãe não pode ser medido no sequenciamento do plasma, pois os *reads* provenientes do feto serão somados aos *reads* da mãe.

Os haplótipos da lista do plasma que possuem frequências relativas menores podem ser provenientes de erros de sequenciamento (conforme foi observado no sequenciamento da mãe e do suposto pai) ou da fração fetal, que espera-se que seja de aproximadamente 10%. Portanto, com o objetivo de eliminar essas informações que provavelmente são provenientes de erros de sequenciamento, desenvolvemos as evidências de paternidade. Utilizamos as evidências para determinar se o conjunto de haplótipos e genótipos observados no *locus* indicam a existência de informação fetal, baseado no haplótipo que o feto herda apenas do pai.

Primeiramente, analisamos a lista de haplótipos proveniente do arquivo do plasma. Com base no artigo *Non-invasive prenatal diagnosis of beta-thalassemia by semiconductor sequencing: a feasibility study in the sardinian population* [9], foi utilizado um intervalo de frequências relativas para determinar os haplótipos que poderiam ser provenientes da fração fetal. O intervalo utilizado na literatura foi de 1,4% a 11%. Dessa forma, separamos dos arquivos do plasma os haplótipos que possuem frequências relativas dentro desse intervalo.

Existem diversas situações que podem ocorrer na herança genética. O feto pode herdar do pai um haplótipo igual ao da mãe. Ou pode herdar um haplótipo diferente do da mãe. A mãe e o suposto pai podem possuir ou não haplótipos em comum. Sendo assim, criamos regras para as evidências de paternidade, onde temos que EV_m é a evidência da paternidade para um dado micro-haplótipo m . Na Tabela 5.1 temos as condições para a atribuição de seu valores.

Em um dado micro-haplótipo m , composto por s SNPs, temos que SP_m é o conjunto de haplótipos do suposto pai, M_m é o conjunto de haplótipos da mãe, P_m é o conjunto de haplótipos do plasma que estejam em uma frequência relativa entre 1,4% e 11% e que sejam diferentes dos haplótipos da Mãe ($P_m \cap M_m = 0$). Do conjunto do plasma, temos

FF_m que é o conjunto de haplótipos do plasma que sejam iguais aos haplótipos do suposto pai ($FF_m = P_m \cap SP_m$) e ES_m é o conjunto de haplótipos do plasma que sejam diferentes dos haplótipos da mãe e do suposto pai ($ES_m = P_m - FF_m$ e $ES_m \cap M_m \cap SP_m = 0$).

Tabela 5.1 - Lista de regras das evidências de paternidade. Dado um micro-haplótipo m , analisamos os haplótipos encontrados e determinamos se existe ou não indícios para o suposto pai ser o pai verdadeiro. Temos que $|SP_m \cap M_m|$ é a quantidade de haplótipos em comum entre o suposto pai e a mãe, $|FF_m|$ é a quantidade de haplótipos no plasma que são iguais aos haplótipos do suposto pai e diferentes dos haplótipos da mãe, $|ES_m|$ é a quantidade de haplótipos do plasma e que não são iguais aos haplótipos do suposto pai e da mãe. D_m^H é o máximo das mínimas distâncias de Hamming. E EV_m é o valor atribuído à evidência de paternidade.

$ SP_m \cap M_m $	$ FF_m $	$ ES_m $	D_m^H	EV_m	Significado
0	0	0	-	0	Não pode ser pai
0	0	≥ 1	$\leq d$	0	Não pode ser pai
0	0	≥ 1	$> d$	0	Não pode ser pai
0	1	0	-	1	Pode ser pai
0	1	≥ 1	$\leq d$	1	Pode ser pai
0	1	≥ 1	$> d$	0	Não pode ser pai
0	2	0	-	1	Pode ser pai
0	2	≥ 1	$\leq d$	1	Pode ser pai
0	2	≥ 1	$> d$	0	Não pode ser pai
1	0	0	-	0,5	Pode ser pai
1	0	≥ 1	$\leq d$	0,5	Pode ser pai
1	0	≥ 1	$> d$	0	Não pode ser pai
1	1	0	-	1	Pode ser pai
1	1	≥ 1	$\leq d$	1	Pode ser pai
1	1	≥ 1	$> d$	0	Não pode ser pai
2	0	0	-	0,5	Pode ser pai
2	0	≥ 1	$\leq d$	0,5	Pode ser pai
2	0	≥ 1	$> d$	0	Não pode ser pai

A distância utilizada nas evidências de paternidade é a distância de Hamming. Dado que existem duas *strings* de mesmo tamanho, a distância de Hamming é definida como sendo o número de posições onde elas diferem entre si [39]. Para um dado micro-haplótipo m composto por s SNPs, temos que $M_m \cup FF_m = \{\alpha^1, \alpha^2, \dots, \alpha^k\}$ onde $\alpha^i = a_1^i a_2^i \dots a_s^i \forall i$ e

$ES_m = \{\beta^1, \beta^2, \dots, \beta^l\}$ onde $\beta^i = b_1^i b_2^i \dots b_s^i \forall i$. Definimos que a distância de Hamming, d_m^H , é dada pela equação 5.1:

$$d_m^H(\alpha, \beta) = \sum_{j=1}^s I(a_j, b_j), \text{ onde } \begin{cases} I = 1, \text{ se } a_j \neq b_j. \\ I = 0, \text{ se } a_j = b_j. \end{cases} \quad (5.1)$$

E temos que D_m^H é a maior distância entre as menores de todas as combinações de $d_m^H(\alpha, \beta)$, conforme pode ser visto na equação 5.2. Isso significa que estamos buscando a maior distância de Hamming de um erro de sequenciamento. Valores muito grandes, podem indicar que não é um erro de sequenciamento, e sim, o haplótipo do pai verdadeiro. Note na Tabela 5.1 que D_m^H depende de um valor d , e iremos ajustar esse parâmetro aos dados dos trios analisados, onde queremos o menor valor de d aceito para erros de sequenciamento.

$$D_m^H = \max_{\beta \in ES_m} \left\{ \min_{\alpha \in (M_m \cup FF_m)} d_m^H(\alpha, \beta) \right\} \quad (5.2)$$

5.1.2 Probabilidade baseada nas evidências de paternidade

Para calcular a probabilidade de paternidade baseada nas evidências, tornamos o valor das evidências em probabilidade. Buscamos um modelo onde a probabilidade baseada nas evidências assuma um valor próximo de 100% para valores de $EV_m = 1$, e próximo de 0% para valores de $EV_m = 0$. Dessa forma, utilizamos a seguinte equação para calcular a probabilidade das evidências, $P(EV_m)$, para um dado micro-haplótipo m , onde os parâmetros γ e δ serão estimados de acordo com os dados.

$$P(EV_m) = \frac{e^{\gamma + EV_m \times \delta}}{1 + e^{\gamma + EV_m \times \delta}} \quad (5.3)$$

Essa é a probabilidade obtida a partir de um micro-haplótipo. Como todos os *loci* analisados são independentes entre si, podemos calcular a probabilidade cumulativa das evidências de paternidade para um número n de micro-haplótipos, P_C^E .

$$P_C^E = \prod_{i=1}^n P(EV_m) \quad (5.4)$$

Nas nossas análises, utilizamos um n de até 20, pois estamos analisando 20 micro-haplótipos.

5.1.3 Probabilidade de paternidade clássica

O cálculo da probabilidade de paternidade já está muito bem estabelecido. Dado que estamos analisando regiões que não possuem alelos nulos e com taxas de mutação desprezíveis. Para realizar tal cálculo, precisamos utilizar o índice de Paternidade. Este é um cálculo estatístico que utiliza as informações genéticas encontradas no sequenciamento. E após calcular o índice de paternidade de cada *locus* analisado, podemos calcular a probabilidade de paternidade.

Nos casos onde as evidências de paternidade indicam que existem informações no plasma que contribuem para o suposto pai ser o pai, ou seja, quando $EV_m \neq 0$, calculamos o índice de paternidade conforme descrito a seguir. Quando as evidências de paternidade não indicam ($EV_m = 0$), consideramos que pode ter ocorrido uma mutação na herança genética. Nesse caso, o índice de paternidade será calculado como uma mutação.

Índice de paternidade

O matemático G. H. Hardy e o médico W. Weinberg são os mais conhecidos por provarem que frequências de alelos e genótipos que estejam em equilíbrio não mudam. Isso assumindo populações infinitas, sem seleção, mutação ou migração e cruzamentos aleatórios [17]

Além disso, usando o equilíbrio de ligação, podemos assumir que alelos em *loci* diferentes são associados aleatoriamente [40]. Com isso, podemos considerar que os micro-haplótipos não possuem dependência um do outro na herança genética. Pois estão localizados há uma distância muito grande um do outro.

Portanto, temos que, desde que os haplótipos sejam herdados independentemente, o equilíbrio de Hardy-Weinberg será aplicado dentro de cada micro-haplótipo e o equilíbrio de ligação será aplicado entre micro-haplótipos [17].

Dado que k é o número de haplótipos de um micro-haplótipo m , temos que o número

total de genótipos possíveis $g_m(k)$ gerados por esse micro-haplótipo é dado por [17]:

$$g_m(k) = \binom{k}{2} + k = \frac{(k^2 + k)}{2} \quad (5.5)$$

Como todos os micro-haplótipos são independentes, podemos fazer o cálculo de $g_m(k)$ para cada micro-haplótipo selecionado. Com os valores de $g_m(k)$ podemos calcular o número de genótipos, G , possivelmente observados para um número n de micro-haplótipos analisados:

$$G = \prod_1^n g_m(k) \quad (5.6)$$

Calculamos os valores $g_m(k)$ dos 20 micro-haplótipos selecionados. Com esses valores, obtivemos G para $n = 20$, e temos que $9,96 \times 10^{41}$ genótipos podem ser observados.

Isso mostra que a quantidade de genótipos formados pelos 20 micro-haplótipos é maior do que a quantidade de indivíduos existentes no mundo. Como não temos as frequências genotípicas, usamos a frequência dos haplótipos de cada micro-haplótipo, levando em conta o equilíbrio de Hardy-Weinberg, o equilíbrio de ligação e a razão da verossimilhança [17].

Dado um micro-haplótipo m , podemos calcular o seu índice de paternidade (IP_m), também denotado como a razão da verossimilhança. Seja X_m a probabilidade dos genótipos observados, em um micro-haplótipo m , dado que o trio de amostras que está sendo testado é verdadeiro, e Y_m a probabilidade dos genótipos, do mesmo micro-haplótipo m , dado que o trio de amostras é falso [40], temos que IP_m é dado por:

$$IP_m = \frac{X_m}{Y_m} \quad (5.7)$$

Vamos usar um exemplo para ilustrar o cálculo do IP_m . Supondo uma mãe com o genótipo BD , um filho com o genótipo AB e um suposto pai com um genótipo AC . Sendo $p(BD)$ a probabilidade do genótipo BD e $p(AC)$ a probabilidade do genótipo AC , $p(A|BD)$ a probabilidade da mãe doar um A dado que o seu genótipo é BD , $p(B|AC)$ a probabilidade do suposto pai doar um B dado que o seu genótipo é AC , $p(A|AC)$ a probabilidade do suposto pai doar um A dado que o seu genótipo é AC e $p(B|BD)$ a

probabilidade da mãe doar um B dado que o seu genótipo é BD , temos que X_m é dado por [40]:

$$\begin{aligned} X_m &= p(BD) \times p(AC) \times [(p(A|BD) \times p(B|AC)) + (p(A|AC) \times p(B|BD))] \\ &= p(BD) \times p(AC) \times [(0 \times 0) + (0,5 \times 0,5)] \\ &= p(BD) \times p(AC) \times 0,25 \end{aligned} \quad (5.8)$$

Para a probabilidade de um pai aleatório PA temos que $p(B|PA)$ é a probabilidade de um pai aleatório doar um B , o que é dado por b (frequência relativa de B na população), e $p(A|PA)$ é a probabilidade de um pai aleatório doar um A , o que é dado por a (frequência relativa de A na população), temos então que Y_m é dado por [40]:

$$\begin{aligned} Y_m &= p(BD) \times p(AC) \times [(p(A|BD) \times p(B|PA)) + (p(A|PA) \times p(B|BD))] \\ &= p(BD) \times p(AC) \times [(0 \times b) + (a \times 0,5)] \\ &= p(BD) \times p(AC) \times 0,5a \end{aligned} \quad (5.9)$$

Portanto, pela equação 5.7 temos que IP_m é dado por:

$$IP_m = \frac{p(BD) \times p(AC) \times 0,25}{p(BD) \times p(AC) \times 0,5a} = \frac{1}{2a} \quad (5.10)$$

Temos no Apêndice C o valor de IP_m para todas as possibilidades de combinação de genótipos dos trios: suposto pai, mãe e filho [40].

Mutação

Quando o filho herda o código genético do pai e da mãe, podem ocorrer mutações no DNA que está sendo herdado. Dessa forma, considerando que isso pode ocorrer nos SNPs que estamos analisando, podemos considerar que os haplótipos observados no feto não serão iguais aos encontrados no suposto pai. Dado um micro-haplótipo m , e sendo

os haplótipos do suposto pai dados por I e J , e H o haplótipo que o feto herdou do pai verdadeiro, $t_{h,i}$ a taxa de mutação do haplótipo I mutar para o haplótipo H , $t_{h,j}$ a taxa de mutação do haplótipo J mutar para o haplótipo H e f_h a frequência populacional do haplótipo H , temos que o índice de paternidade nos casos de mutação é dado pela seguinte equação [40]:

$$IP_m = \frac{t_{h,i} + t_{h,j}}{2f_h} \quad (5.11)$$

No entanto, a taxa de mutação de apenas um SNP é da ordem de 10^{-8} . Se dois SNPs sofrerem mutação, essa taxa será da ordem de 10^{-16} , e assim por diante. E a menor frequência que um haplótipo tem no banco do *1000 Genomes* é de 0,0002. Portanto, o maior valor que IP_m poderá assumir nos casos de mutação será da ordem de 5×10^{-5} . Para simplificar os cálculos nos casos onde a evidência de paternidade indicar que pode ter ocorrido uma mutação ($EV_m = 0$), vamos assumir nesses casos que $IP_m = 5 \times 10^{-5}$.

Probabilidade de paternidade

Como os micro-haplótipos são independentes, temos que o índice de paternidade cumulativo (IP_C) é dado pela multiplicação de todos os IP_m [40]. Para um número n de micro-haplótipos, temos:

$$IP_C = \prod_{i=1}^n IP_m \quad (5.12)$$

O IP_C é uma medida da força da evidência genética. Essa evidência pode se adequar à hipótese de que o suposto pai é o pai biológico ou à hipótese de que outra pessoa é o pai biológico. Com isso, o IP_C pode variar de zero a infinito (ver Tabela 5.2) [40].

Tabela 5.2 - Valores do IP_C e seus significados.

Valor do IP_C	Significado
$IP_C = 0$	Exclusão da paternidade
$0 < IP_C < 1$	Exclusão da paternidade > Inclusão da paternidade
$IP_C = 1$	Não conclusivo
Continua na próxima página...	

Tabela 5.2 - Continuação

Valor do IP_C	Significado
$IP_C > 1$	Inclusão da paternidade (quanto maior, maior a evidência)

Sendo P a probabilidade a priori do suposto pai ser o pai biológico, temos que $(1 - P)$ é a probabilidade de outro indivíduo ser o pai [40]. Portanto, a probabilidade de paternidade é dada por:

$$\text{Probabilidade de paternidade} = \frac{IP_C \times P}{IP_C \times P + (1 - P)} \quad (5.13)$$

Assumindo uma probabilidade a priori de $P = 0,5$, temos então que a probabilidade de paternidade posterior (W), ou seja, a probabilidade de paternidade dada pela equação 5.13, é dada por:

$$W = \frac{IP_C}{IP_C + 1} \quad (5.14)$$

5.2 Resultados e Discussão

5.2.1 Obtenção da fração fetal

O intervalo utilizado na literatura para detectar frações fetais dos haplótipos, que o feto herda apenas do pai, é de 1,4% a 11% [9]. Para verificar se as possíveis frações fetais dos trios analisados estavam dentro desse intervalo, utilizamos os trios onde o suposto pai foi determinado como pai biológico. Nesses trios, selecionamos da lista de haplótipos encontrados do arquivo do plasma (após a checagem da qualidade, ver no Capítulo 4) os haplótipos que o feto pode ter herdado apenas do pai. E fizemos um histograma das frequências relativas dos possíveis haplótipos que o feto herdou do pai, ver Figura 5.1.

Observamos na Figura 5.1 que a maior concentração dos possíveis haplótipos fetais está no intervalo de 1,4% a 11%, conforme esperado. Também podemos observar que

existem haplótipos fetais com frequências relativas muito acima do esperado, acima de 12,5%, o que resultaria em uma fração fetal acima de 25%. Provavelmente estes sejam oriundos de erros de sequenciamento acima do normal em um micro-haplótipo ou de uma amostra com problemas no sequenciamento do plasma. No Capítulo 4, observamos que os erros de sequenciamento também estão no mesmo intervalo de frequências relativas que os haplótipos fetais herdados do pai (abaixo de 10%). No entanto, frações fetais muito baixas não podem ser aceitas. Portanto, levando em conta os dados observados no gráfico da Figura 5.1, e buscando obter informação confiável dos possíveis haplótipos fetais herdados do pai, alteramos o intervalo de detecção dos mesmos para 1% a 12%.

Figura 5.1 - Histograma da frequência relativa dos possíveis haplótipos que o feto herdou apenas do pai. As frequências relativas esperadas para os haplótipos que o feto herda apenas do pai são valores abaixo 11%. No entanto, observamos alguns valores muito acima de 12,5%.

Sabemos que a presença do DNA fetal pode influenciar no desbalanço das frequências relativas dos haplótipos da mãe no sequenciamento do plasma. No entanto, esses haplótipos ainda devem estar presentes no plasma. Portanto, estabelecemos que no sequenciamento do plasma, os haplótipos da mãe deveriam estar presentes e com frequências relativas maiores do que 12%. Estabelecemos também que, com exceção dos haplótipos da mãe, mais nenhum outro haplótipo poderia ter frequência relativa maior do que 12%. Utilizamos estas condições para tentar eliminar regiões cujo sequenciamento tenha gerado muitos erros. Ao fazer novamente o histograma das frequências relativas dos possíveis haplótipos fetais, após aplicar as condições acima, obtivemos o gráfico da Figura 5.2.

Observamos que os possíveis haplótipos que o feto herdou do pai estavam restritos ao intervalo de 1% a 12%, conforme esperado após as restrições impostas. Como não existe uma quantidade mínima de micro-haplótipos onde devemos detectar a fração fetal,

não podemos impor um número mínimo de micro-haplótipos em que os haplótipos do suposto pai serão diferentes dos haplótipos da mãe. No entanto, esperamos que a frequência relativa dos haplótipos fetais herdados apenas do pai sejam próximos entre diferentes micro-haplótipos. Dessa forma, fizemos um *barplot* da frequência relativa dos possíveis haplótipos herdados do pai, para os 20 micro-haplótipos analisados (Figura 5.3). Utilizamos somente as amostras onde sabemos que o suposto pai testado foi determinado como pai verdadeiro.

Figura 5.2 - Histograma da frequência relativa dos possíveis haplótipos que o feto herdou do pai após aplicar novas condições. Estabelecemos que os possíveis haplótipos que o feto herda do pai estariam entre 1% e 12% de frequência relativa, e nenhum outro haplótipo, além dos haplótipos da mãe, poderiam ter frequências relativas acima de 12%. Dessa forma, eliminamos os erros de sequenciamento massivos. Observamos que após aplicar estas condições, as frequências relativas dos possíveis haplótipos que o feto herda do pai estão dentro das faixas esperadas (entre 1% e 12%).

Podemos observar pela Figura 5.3 que a frequência relativa dos possíveis haplótipos herdados do pai não se manteve constante entre os micro-haplótipos da mesma amostra, por exemplo, as amostras T11, T35 e T56. Isso significa que não podemos garantir que todos os possíveis haplótipos fetais herdados do pai estão sendo detectados. E conforme visto no gráfico da Figura 5.1, existem haplótipos que possuem frequências relativas abaixo do intervalo de confiança (entre 1% e 12%). Nas próximas etapas, iremos analisar o impacto desse intervalo no cálculo da probabilidade de paternidade.

Figura 5.3 - Barplot da frequência relativa dos possíveis haplótipos herdados do pai, nos 20 micro-haplótipos utilizados nas análises. Utilizamos apenas as amostras onde o suposto pai foi determinado como pai verdadeiro, e que possuíam pelo menos um micro-haplótipo onde o possível haplótipo herdado do pai foi detectado. Observamos que, para uma mesma amostra (coluna), os valores das frequências relativas não se manteve constante entre os micro-haplótipos.

5.2.2 Evidências de paternidade

Nos Materiais e Métodos, criamos as evidências de paternidade (EV_m) e precisamos agora definir e estimar alguns parâmetros. Primeiramente, precisamos determinar d , que

é a distância mínima aceita para um erro de sequenciamento. Esperamos que nos casos onde o suposto pai é o pai verdadeiro, D_m^H seja pequeno, pois esperamos que os erros de sequenciamento sejam pequenos. E nos casos onde o suposto pai não é o pai verdadeiro, esperamos que D_m^H seja grande, pois isso indicaria que o que estamos definindo como erro de sequenciamento é na verdade o haplótipo que o feto herdou do pai verdadeiro.

Utilizamos o novo intervalo de detecção dos possíveis haplótipos que o feto herda do pai (1% a 12%) e separamos as amostras entre trios onde o suposto pai é o pai verdadeiro (*Verdadeiro*) e trios onde o suposto pai não é o pai verdadeiro (*Falso*). Utilizando o grupo *Verdadeiro*, extraímos para cada $\beta \in ES_m$ o menor $d_m^H(\alpha, \beta)$, com $\alpha \in (M_m \cup FF_m)$ (dado pela equação 5.1). Assim, para cada haplótipo do erro de sequenciamento, temos a menor distância de Hamming. E fazendo o mesmo para o grupo *Falso*, esperamos obter no grupo *Falso* valores mais altos do que no grupo *Verdadeiro*. Fizemos então um histograma dos valores de distância obtidos para os dois casos, e utilizando valores de qualidade das bases diferentes, ver Figura 5.4.

Figura 5.4 - Histograma da distância de Hamming dos erros de sequenciamento. Os dados foram separados em dois grupos, *Falso* e *Verdadeiro*. No grupo *Falso*, o suposto pai não é o pai verdadeiro e no grupo *Verdadeiro*, o suposto pai é o pai verdadeiro. A distância de Hamming dos erros de sequenciamento também foi obtida para diferentes valores da qualidade das bases dos sequenciamentos (Q20, Q25 e Q30). Observamos que os valores das distâncias de Hamming para o grupo *Verdadeiro* e para o grupo *Falso* são semelhantes.

Podemos observar na Figura 5.4 que tanto nos dados onde o suposto pai é o pai verdadeiro e nos dados onde o suposto pai não é o pai verdadeiro, d_m^H dos erros de sequenciamento é alta (valores entre 10 e 11). E mesmo utilizando valores diferentes da qualidade das bases,

essa distância se manteve. Com isso, podemos concluir que os erros de sequenciamento não podem ser utilizados para determinar se existe ou não evidência da paternidade em um *locus*. Dessa forma, utilizamos uma nova abordagem para determinar os valores de EV_m (ver Tabela 5.3).

Tabela 5.3 - Lista das novas regras das evidências de paternidade. Dado um micro-haplótipo m , analisamos os haplótipos encontrados e determinamos se existe ou não indícios para o suposto pai ser o pai verdadeiro. Temos que $|SP_m \cap M_m|$ é a quantidade de haplótipos em comum entre o suposto pai e a mãe, $|FF_m|$ é a quantidade de haplótipos no plasma que são iguais ao haplótipo do suposto pai e diferentes dos haplótipos da mãe (estes se encontram no intervalo de 1% e 12% no plasma). EV_m é o valor atribuído à evidência de paternidade.

$ SP_m \cap M_m $	$ FF_m $	EV_m	Significado	$ SP_m \cap M_m $	$ FF_m $	EV_m	Significado
0	0	0	Não pode ser pai	1	0	0,5	Pode ser pai
0	1	1	Não pode ser pai	1	1	1	Pode ser pai
0	2	1	Pode ser pai	2	0	0,5	Pode ser pai

Obtivemos então os valores de EV_m para os 20 micro-haplótipos dos 60 trios analisados. Nos casos onde não existia informação com qualidade no micro-haplótipo, $EV_m = \text{NA}$. Para analisar melhor esses trios, fizemos a Tabela 5.4, para cada micro-haplótipo com qualidade, ou seja, $EV_m \neq \text{NA}$, atribuímos as regras apresentadas na Tabela 5.3.

Tabela 5.4 - Lista de qualidade dos trios. Onde temos o trio analisado (Trio), o resultado obtido pela empresa Genomic (exclusão ou inclusão), a quantidade total de micro-haplótipos com qualidade (Total). $EV(1)$ indica o número de micro-haplótipos com $EV_m = 1$, $EV(0,5)$ indica o número de micro-haplótipos com $EV_m = 0,5$ e $EV(0)$ indica o número de micro-haplótipos com $EV_m = 0$.

Trio	Genomic	Total	$EV(1)$	$EV(0,5)$	$EV(0)$	Trio	Genomic	Total	$EV(1)$	$EV(0,5)$	$EV(0)$
T01	Exclusão	0	0	0	0	T02	Inclusão	0	0	0	0
T03	Inclusão	1	0	1	0	T04	Exclusão	2	0	2	0
T05	Inclusão	0	0	0	0	T06	Exclusão	0	0	0	0
T07	Exclusão	0	0	0	0	T08	Inclusão	1	0	1	0
T09	Exclusão	17	4	10	3	T10	Inclusão	2	0	1	1

Continua na próxima página...

Tabela 5.4 - Continuação

Trio	Genomic	Total	$EV(1)$	$EV(0,5)$	$EV(0)$	Trio	Genomic	Total	$EV(1)$	$EV(0,5)$	$EV(0)$
T11	Inclusão	15	6	8	1	T12	Exclusão	15	4	10	1
T13	Inclusão	1	0	0	1	T14	Inclusão	8	2	5	1
T15	Inclusão	7	3	4	0	T16	Inclusão	12	3	9	0
T17	Inclusão	5	1	4	0	T18	Inclusão	12	3	8	1
T19	Exclusão	2	0	1	1	T20	Inclusão	1	1	0	0
T21	Inclusão	1	1	0	0	T22	Exclusão	0	0	0	0
T23	Inclusão	0	0	0	0	T24	Inclusão	0	0	0	0
T25	Exclusão	0	0	0	0	T26	Exclusão	0	0	0	0
T27	Exclusão	0	0	0	0	T28	Inclusão	0	0	0	0
T29	Inclusão	0	0	0	0	T30	Inclusão	0	0	0	0
T31	Inclusão	11	3	7	1	T32	Exclusão	10	3	6	1
T33	Inclusão	14	6	7	1	T34	Inclusão	14	5	9	0
T35	Inclusão	14	7	7	0	T36	Exclusão	0	0	0	0
T37	Inclusão	14	5	9	0	T38	Inclusão	8	2	6	0
T39	Exclusão	15	6	5	4	T40	Exclusão	16	2	13	1
T41	Inclusão	15	4	10	1	T42	Inclusão	15	7	7	1
T43	Exclusão	18	2	12	4	T44	Exclusão	17	2	13	2
T45	Exclusão	12	1	9	2	T46	Exclusão	14	1	11	2
T47	Inclusão	13	7	5	1	T48	Exclusão	12	0	9	3
T49	Exclusão	14	0	12	2	T50	Inclusão	14	2	12	0
T51	Exclusão	15	4	7	4	T52	Exclusão	15	2	11	2
T53	Inclusão	17	6	11	0	T54	Exclusão	16	3	9	4
T55	Inclusão	16	6	10	0	T56	Inclusão	14	5	8	1
T57	Inclusão	15	7	8	0	T58	Inclusão	17	7	10	0
T59	Inclusão	14	9	5	0	T60	Exclusão	15	3	9	3

Na Capítulo 4, demonstramos que conforme utilizamos mais micro-haplótipos nas análises, a precisão do cálculo da probabilidade de paternidade aumenta, ou seja, se torna mais próxima de 100%. Também observamos que com o aumento da quantidade de micro-haplótipos, a chance de ocorrer falsos positivos diminui. Levando isso em conta, iremos analisar os trios com 12 ou mais micro-haplótipos, pois as chances de ocorrerem falsos positivos são de 11 em 52.000 simulações.

Dos 60 trios de amostras analisados, 31 deles possuíam 12 ou mais micro-haplótipos com qualidade. Analisando os casos onde o suposto pai é o pai verdadeiro, esperamos que nenhum deles possuam micro-haplótipo com $EV_m = 0$. No entanto, tivemos sete casos de inclusão com um *locus* indicando exclusão (T11, T18, T33, T41, T42, T47 e

T56). Agora analisando os casos de exclusão, observamos que dois deles possuem apenas um *locus* indicando a exclusão (T12 e T40) e os outros 12 casos possuem dois ou mais micro-haplótipos onde $EV_m = 0$.

Buscamos descobrir os motivos porque os casos onde o suposto pai é o pai verdadeiro possuíam um *locus* indicando exclusão. Os resultados estão apresentados na Tabela 5.5.

Tabela 5.5 - Lista dos motivos porque os resultados estavam inconsistentes. N° Micro indica a quantidade de micro-haplótipos que o trio possui com qualidade para ser analisado. Locus indica o micro-haplótipo que está indicando exclusão. E problemas encontrados são os motivos que estão levando à exclusão da paternidade no Trio analisado.

Trio	N° Micro	Locus	Problemas encontrados
T11	15	M04	Não existe haplótipo do SP no sequenciamento do PL.
T18	12	M04	Haplótipo do SP se encontra numa fração relativa de 0,47%.
T33	14	M08	Não existe haplótipo do SP no sequenciamento do PL.
T41	15	M04	Haplótipos do SP se encontram numa fração relativa de 0,98% e 0,07%.
T42	15	M11	Haplótipo do SP se encontra numa fração relativa de 0,74%.
T47	13	M05	Haplótipos do SP se encontram numa fração relativa de 0,40% e 0,20%.
T56	14	M03	Haplótipo do SP se encontra numa fração relativa de 0,01%.

Da Tabela 5.5, podemos observar que os principais motivos para a falsa exclusão de paternidade ocorrer era a não detecção do haplótipo que o feto herdou do pai. Na maioria das situações ocorreu porque estava abaixo do limite mínimo de detecção de 1%. E em alguns casos, ocorreu porque o haplótipo do pai não foi encontrado (T11 e T33).

Observamos que existem sete casos de falsos negativos e dois casos de verdadeiros negativos que se comportam da mesma forma que os falsos negativos, apenas um *locus* indicando exclusão. Concluímos que a evidência para esses nove casos serem considerados exclusão ou inclusão é muito fraca. Portanto, para termos casos de inclusão da paternidade, não podem existir micro-haplótipos que indiquem a exclusão ($EV_m = 0$). Para os casos de exclusão, devem existir dois ou mais micro-haplótipos onde $EV_m = 0$. E por fim, se $EV_m = 1$, o exame será inconclusivo.

5.2.3 Probabilidade de paternidade - Evidências de paternidade

Utilizando os dados das evidências de paternidade dos 20 micro-haplótipos de cada um dos 60 trios analisados, obtivemos uma matriz de evidências. Para calcular a probabilidade de paternidade, precisamos calcular as probabilidades das evidências, dada pela equação 5.3. Primeiramente, estimamos os parâmetros γ e δ utilizando o método de estimativa da Máxima Verossimilhança. Usamos a função *nlm* (*non-linear minimization*) do *R* e a matriz de valores das evidências de paternidade. Estimamos que $\gamma = -0.9985615$ e $\delta = 2.161142$. Fizemos também um *bootstrap*, sem reposição, com 30 trios e estimamos os parâmetros γ e δ , esse processo foi repetido 10.000 vezes. Construímos um gráfico com os valores de γ e δ estimado e simulados com o *bootstrap*, ver Figura 5.5.

Figura 5.5 - Dotplot dos valores estimado e simulado de γ e δ . Os valores simulados foram obtidos de 1000 repetições da simulação fazendo *bootstrap* sem reposição de 30 trios. Observamos que o valor estimado se encontra dentro do intervalo de confiança de 95% dos valores simulados.

Para cada *bootstrap* realizados utilizando 30 trios (dos 60 trios de amostras), calculamos γ e δ (Figura 5.5). Desses dados, obtivemos que os intervalos de confiança de 95% desses parâmetros são: $-1.694484 \leq \gamma \leq -0.2319329$ e $1.580257 \leq \delta \leq 2.88681$. Sendo assim, os valores estimados de γ e δ estão dentro no intervalo de confiança de 95% dos valores obtidos das simulações utilizando *bootstrap*.

Como apresentado anteriormente, trios com poucos micro-haplótipos cobertos não possuem qualidade para serem analisados. Por isso, iremos fazer os cálculos apenas para os

trios que possuem 12 ou mais micro-haplótipos cobertos. Os trios que possuem apenas um *locus* onde $EV_m = 0$, foram classificados como *Inconclusivos*, e por isso não iremos realizar os cálculos também. Portanto, utilizando apenas os trios que não possuem indícios de exclusão e os trios cujos indícios de exclusão são maiores (dois ou mais *loci* com $EV_m = 0$), calculamos as probabilidades das evidências, dada pela equação 5.3. Calculando então $P(EV_m)$ para todos os micro-haplótipos de um trio, podemos calcular a probabilidade de paternidade baseada nas evidências de paternidade, dada pela equação 5.4. Na Figura 5.6, apresentamos um gráfico com os valores da probabilidade de paternidade calculados para todos os trios que possuíam as condições impostas. O esperado é que os trios sem indícios de exclusão tenham valores de probabilidade de paternidade próximos de um, e trios que possuem indícios de exclusão, possuam valores de probabilidade de paternidade próximos de zero.

Figura 5.6 - Dotplot de P_C^E dos dados da Genomic. A probabilidade de paternidade baseada nas evidências de paternidade foi calculada para os casos onde haviam indícios de inclusão e exclusão. Casos inconclusivos e casos onde o número de micro-haplótipos com qualidade era menor que 12 não foram calculados. Na legenda Genomic, *Exclusão* indica os trios onde o suposto pai não foi determinado como o pai verdadeiro e *Inclusão* indica os trios onde o suposto pai foi determinado como o pai verdadeiro. O tamanho dos pontos indica a quantidade de micro-haplótipos que foram utilizados nas análises. Observamos que os casos de *Exclusão* estão indicando exclusão, porém, os casos de *Inclusão* não estão sendo determinados corretamente como inclusão.

Do gráfico da Figura 5.6, podemos observar que os trios que são casos de exclusão dados pela Genomic, também foram considerados exclusão pelo nosso método. No entanto, os casos de inclusão não apresentaram valores de probabilidade de paternidade aceitáveis

para indicar a inclusão. Podemos concluir que o método baseado somente nas evidências de paternidade ainda não possui a capacidade de distinguir casos de exclusão e inclusão de paternidade.

5.2.4 Probabilidade de paternidade - Clássica

Utilizando o modelo clássico para calcular a probabilidade de paternidade, precisamos calcular os índices de paternidade (IP_m) de cada micro-haplótipo. Novamente, utilizando apenas os trios que possuem 12 ou mais micro-haplótipos com qualidade e excluindo os casos inconclusivos (apenas um *locus* onde $EV_m = 0$), calculamos os índices de paternidade seguindo as regras apresentadas no Apêndice C. Como não podemos medir qual haplótipo o feto herda da mãe, quando existe a dúvida, calculamos todas as possibilidades de combinações de IP_m e utilizamos o menor valor. Com o valor de IP_m calculado para todos os micro-haplótipos, calculamos o índice de paternidade cumulativo (IP_c), dado pela equação 5.12.

Figura 5.7 - Dotplot de W dos dados da Genomic. A probabilidade de paternidade clássica foi calculada para os casos onde haviam indícios de inclusão e exclusão. Casos inconclusivos e casos onde o número de micro-haplótipos com qualidade era menor que 12 não foram calculados. Na legenda Genomic, *Exclusão* indica os trios onde o suposto pai não foi determinado como o pai verdadeiro e *Inclusão* indica os trios onde o suposto pai foi determinado como o pai verdadeiro. O tamanho dos pontos indica a quantidade de micro-haplótipos que foram utilizados nas análises. Observamos que os casos de *Exclusão* estão indicando exclusão e os casos de *Inclusão* estão indicando inclusão.

Após calcular o IP_c , calculamos a probabilidade de paternidade posterior (W), dada

pela equação 5.14 para todos os trios que possuíam mais de 12 micro-haplótipos com qualidade e não eram casos inconclusivos (apenas um *locus* com $EV_m = 0$). Apresentamos na Figura 5.7 os valores de W calculados para estes trios.

Observamos a partir da Figura 5.7, que os casos de que a Genomic determinou como exclusão, também foram determinados como exclusão pelo nosso método, pois o valor de W está próximo de zero. Nos casos de inclusão, observamos que os valores de W também indicam inclusão no nosso método, pois estão próximos de um. Além disso, o valor de W para todos os trios onde foi indicada a inclusão é superior a 0,99.

Figura 5.8 - Heatmap da simulação suposto pai versus mãe e plasma, mínimo de 12. Nessa simulação, testamos a paternidade do suposto pai versus a dupla mãe e plasma. Nos casos onde o número de micro-haplótipos é menor que 12, definimos como *Menos de 12* e não analisamos porque a qualidade é baixa. Nos casos onde o número de micro-haplótipos é maior ou igual a 12, os dados são analisados. Nesses casos, se existir um micro-haplótipo onde $EV_m = 0$, é classificado como *Inconclusivo*. Se existir mais de um micro-haplótipo onde $EV_m = 0$, é classificado como *Exclusão*, e se não existir micro-haplótipo onde $EV_m = 0$, é classificado como *Inclusão*. Para os casos de *Inclusão*, calculamos a probabilidade de paternidade posterior (W), e esta é proporcional à transparência da cor.

Com o objetivo de analisar se existiriam falsas inclusões de paternidade caso testássemos

a paternidade do suposto pai de um trio contra a mãe e o plasma de outro trio, criamos uma simulação de não pais. Nessa simulação, comparamos todos os supostos pais contra todas as duplas mães e plasma. O esperado são inclusões apenas nas diagonais onde sabemos que o resultado é de inclusão (T16, T34, T35, T37, T50, T53, T55, T57, T58 e T59) e em alguns casos especiais relacionados aos trios.

Figura 5.9 - Heatmap da simulação suposto pai versus mãe e plasma, mínimo de 15. Nessa simulação, testamos a paternidade do suposto pai versus a dupla mãe e plasma. Nos casos onde o número de micro-haplótipos é menor que 15, definimos como *Menos de 15* e não analisamos porque a qualidade é baixa. Nos casos onde o número de micro-haplótipos é maior ou igual a 15, os dados são analisados. Nesses casos, se existir um micro-haplótipo onde $EV_m = 0$, é classificado como *Inconclusivo*. Se existir mais de um micro-haplótipo onde $EV_m = 0$, é classificado como *Exclusão*, e se não existir micro-haplótipo onde $EV_m = 0$, é classificado como *Inclusão*. Para os casos de *Inclusão*, calculamos a probabilidade de paternidade posterior (W), e esta é proporcional à transparência da cor.

Existem alguns trios que são casos especiais de amostras. Alguns deles são repetições de análises, são eles: T05 e T20, T06 e T19, T10 e T13, T11 e T18, T21 e T30, T40 e T44. Dessa forma, é esperado que nos casos de inclusão, o suposto pai de um trio indique inclusão com a dupla mãe e plasma do outro trio. Outro caso especial são os trios T53 e T54. Nesse trio, existiam dois supostos pais, por isso, a mãe e o plasma do trio T53 são

os mesmos do trio T54.

No *heatmap* da Figura 5.8, observamos que existem mais casos de inclusão do que o esperado, por exemplo os trios T16, T37 e T59. Isso significa que estão ocorrendo muitas falsas inclusões de paternidade. Como demonstrado na Capítulo 4, quanto mais micro-haplótipos utilizados nas análises, maior a precisão do cálculo de paternidade e menores as chances de ocorrerem falsos positivos. Além disso, os erros de sequenciamento inerentes da técnica podem gerar haplótipos que tornam a inclusão de um falso pai possível. E no *heatmap* da Figura 5.9, podemos observar que ao exigimos um mínimo de 15 micro-haplótipos sendo analisados, diminuem os casos de falsos positivos.

5.2.5 Validação do método em dados simulados

Podemos concluir através de nossas análises que muitos dos dados obtidos do banco de dados da Genomic estavam com a qualidade abaixo do necessário para que os cálculos e testes pudessem ser realizados, principalmente os dados do plasma. Para analisar se o método funcionaria com dados reais que estivessem com a qualidade mínima necessária, criamos algumas simulações usando dados reais. Utilizamos para as simulações os trios onde o sequenciamento do suposto pai e da mãe estavam com qualidade boa, ou seja, ambos com mais de 12 micro-haplótipos cobertos. Dos trios que se encaixavam em casos especiais (repetições de análises), escolhemos apenas um deles para ser utilizado nas simulações. Dessa forma, obtivemos 36 pares de amostras suposto pai e mãe: T09, T10, T11, T12, T14, T15, T16, T17, T31, T32, T33, T34, T35, T37, T38, T39, T41, T42, T43, T44, T45, T46, T47, T48, T49, T50, T51, T52, T53, T54, T55, T56, T57, T58, T59 e T60.

Primeiramente, fizemos uma simulação variando a fração fetal. O objetivo era analisar a influência da fração fetal na porcentagem de casos de *Inclusão*, *Exclusão*, *Inconclusivos* e *Menos de 12* micro-haplótipos para serem analisados. Nessa simulação, utilizamos os dados do suposto pai e da mãe para criar um arquivo de plasma artificial. No arquivo artificial, a porcentagem de um dos haplótipos do pai foi adicionada artificialmente de acordo com a fração fetal simulada. Variamos a fração fetal de 1% a 25%, no intervalo de 1% a 4% e de 22% a 25%, a fração foi acrescida de 0,1%, e no intervalo de 4% a 22%, a fração foi acrescida de 1%. Para cada fração fetal, a simulação foi repetida 100 vezes. Os

resultados foram separados em quatro casos, *Menos de 12* são os casos onde não haviam 12 ou mais micro-haplótipos para serem analisados. *Inconclusivo* são os casos onde existem 12 ou mais micro-haplótipos para serem analisados, porém, existe um *locus* onde $EV_m = 0$. *Inclusão* são os casos onde existem 12 ou mais micro-haplótipos para serem analisados e todos eles indicam a inclusão da paternidade. E *Exclusão* são os casos onde existem 12 ou mais micro-haplótipos para serem analisados, porém, dois ou mais deles indicam exclusão de paternidade ($EV_m = 0$). O *boxplot* do resultado da simulação utilizando os 36 pares de amostras suposto pai e mãe está apresentado na Figura 5.10.

Figura 5.10 - Boxplot da simulação da fração fetal utilizando dados da Genomic. A fração fetal foi acrescida de 0,1% entre 1% e 4%, acrescida de 1% entre 4% e 22%, e por fim, acrescida de 0,1% entre 22% e 25%. Para um par suposto pai e mãe, a simulação foi realizada 100 vezes em cada valor de fração fetal. O gráfico contém o resultado da simulação realizada nos 36 pares de amostras suposto pai e mãe obtidos pela Genomic.

Observando os resultados da Figura 5.10, podemos concluir que o *script* é capaz de detectar frações fetais próximas a 2%, e a porcentagem de inclusões foram altas, mesmo para esses valores baixos de fração fetal. Observamos também que conforme a fração fetal atinge valores próximos de 22%, a porcentagem de *Inclusão* começa a cair e a porcentagem de *Menos de 12* aumenta. Isso é esperado, pois o *script* foi desenvolvido para detectar haplótipos, herdados apenas do pai, no intervalo de 1% a 12%, o que significa uma fração fetal de 2% a 24%.

Nos resultados apresentados na Figura 5.10, observamos que existem alguns *outliers* nas simulações. Por isso, fizemos a Tabela 5.6, onde temos a quantidade de micro-haplótipos cobertos no suposto pai, na mãe e a quantidade em comum entre as duas amostras ($SP \cap M$).

Tabela 5.6 - Lista da quantidade de micro-haplótipos cobertos em cada dupla e por amostra: suposto pai (SP) e mãe (M). Apresentamos também a quantidade de micro-haplótipos que estão sendo cobertos em comum às duas amostras analisadas ($SP \cap M$).

Trio	SP	M	$SP \cap M$	Trio	SP	M	$SP \cap M$	Trio	SP	M	$SP \cap M$
T09	19	18	18	T10	17	19	17	T11	18	17	17
T12	16	18	16	T14	16	15	14	T15	14	15	12
T16	16	16	16	T17	16	17	15	T31	18	17	16
T32	19	18	18	T33	16	18	16	T34	18	19	18
T35	19	18	18	T37	19	19	19	T38	16	13	12
T39	16	18	16	T41	17	16	15	T42	18	17	17
T43	18	19	18	T44	19	18	18	T45	15	13	13
T46	14	16	14	T47	15	13	13	T48	18	14	14
T49	17	15	14	T50	17	15	14	T51	18	15	15
T52	16	17	15	T53	19	17	17	T54	16	17	16
T55	17	18	17	T56	17	16	15	T57	17	18	16
T58	17	18	17	T59	18	18	17	T60	17	18	16

Apesar de controlarmos a porcentagem da fração fetal dos dados do suposto pai e da mãe nas simulações, nós não excluimos os erros de sequenciamento que estavam acontecendo nas amostras. Por isso, ocorreram os *outliers*.

Na nossa segunda simulação, criamos um *dataset* de 60 trios, onde 35 eram casos de inclusão e 25 eram casos de exclusão. Utilizamos as 36 amostras já citadas para criar esse conjunto. Nas nossas simulações, criamos um conjunto onde a fração fetal é de 10%, supondo que a coleta das amostras seria em torno de 10 semanas de gravidez e a fração fetal esperada nesse tempo de gestação é de 10%.

Utilizando então uma fração fetal simulada de 10%, criamos 36 trios de suposto pai, mãe e plasma simulado. Obtivemos desses 36 trios a frequência relativa dos possíveis haplótipos fetais herdados do pai, e criamos um *barplot* com essas frequências (Figura 5.11). O esperado dessa simulação são frequências relativas próximas de 5% ,metade da

fração fetal é herdada do pai e a simulação foi realizada com uma fração fetal de 10%.

Podemos observar no gráfico de *barplot* da Figura 5.11, que conforme esperado, as

frequência relativa dos possíveis haplótipos herdados do pai estavam em torno de 5%. Utilizamos 35 dos trios para formarem o grupo de *Inclusão*, onde o suposto pai é o pai verdadeiro. Para formar o grupo de *Exclusão*, onde o suposto pai não é o pai verdadeiro, utilizamos o 36º suposto pai, combinado com outras 25 mães para criar o plasma artificial. E novamente comparamos o suposto pai correspondente à mãe e o plasma artificial. Obtivemos dessa comparação uma matriz de evidências de paternidade (EV_m) dos 60 trios e estimamos os valores de γ e δ utilizando o método de estimativa da Máxima Verossimilhança com a função *nlm* do *R*. Os valores estimados foram $\gamma = -1.405046$ e $\delta = 2.779685$. Fizemos também um *bootstrap* com 30 trios e estimamos os parâmetros γ e δ , o *bootstrap* foi repetido 10.000 vezes. Construimos um gráfico com os valores de γ e δ estimado e simulados com o *bootstrap*, ver Figura 5.12.

Figura 5.12 - Dotplot dos valores estimado e simulado de γ e δ dos dados artificiais. Os valores simulados foram obtidos de 10.000 repetições da simulação fazendo *bootstrap* sem reposição de 30 trios. Observamos que o valor estimado se encontra dentro do intervalo de confiança de 95% dos valores simulados.

Os intervalos de confiança de 95% dos valores simulados de γ e δ através do *bootstrap* de 30 trios são: $-1.919682 \leq \gamma \leq -0.8890877$ e $2.344354 \leq \delta \leq 3.312623$. Podemos observar que os valores estimados de δ e γ estão dentro do intervalo de confiança obtido pelo *bootstrap* de 30 trios.

Utilizando os valores de γ e δ estimados pelos dados criados artificialmente, calculamos a probabilidade das evidências, $P(EV_m)$, para todas as amostras que possuíam 12 ou mais micro-haplótipos com qualidade e não fossem casos *Inconclusivos* (onde um micro-

halótipo apenas possui $EV_m = 0$). Com os valores de $P(EV_m)$, calculamos a probabilidade cumulativa das evidências de paternidade (P_C^E), dada pela equação 5.4. Construímos o gráfico da Figura 5.13, onde apresentamos os valores calculados para P_C^E para todos os trios que possuíam indícios de inclusão (nenhum *locus* com $EV_m = 0$) ou exclusão (dois ou mais *loci* com $EV_m = 0$). Esperamos nessa simulação que os casos de inclusão possuam valores de P_C^E próximos de um e nos casos de exclusão, valores próximos de zero.

Figura 5.13 - Dotplot de P_C^E dos dados artificiais. A probabilidade de paternidade baseada nas evidências de paternidade foi calculada para os casos simulados com 12 ou mais micro-haplótipos com qualidade e indícios de ser exclusão ou inclusão. Na legenda Simulação, *Exclusão* indica os trios onde o suposto pai foi simulado como o pai não verdadeiro e *Inclusão* indica os trios onde o suposto pai foi simulado como sendo o pai verdadeiro. O tamanho dos pontos é proporcional à quantidade de micro-haplótipos que foram utilizados nas análises. Observamos que alguns casos de *Exclusão* estão sendo classificados incorretamente (T45_F e T47_F), e todos os casos de *Inclusão* também foram classificados incorretamente.

Como observado anteriormente, mesmo utilizando dados simulados com qualidade superior aos dados reais da Genomic, o modelo de cálculo da probabilidade de paternidade, baseado nas evidências de paternidade, ainda não é capaz de distinguir os casos de exclusão e inclusão de paternidade.

Agora, utilizando o modelo de cálculo da probabilidade clássica, calculamos os índices de paternidade (IP_m) para todos os trios que possuíam 12 ou mais micro-haplótipos com qualidade, e não fossem casos *Inconclusivos* (apenas um *locus* onde $EV_m = 0$). E com os valores de IP_m , calculamos o índice de paternidade cumulativo (IP_C) e a probabilidade de paternidade clássica (W), dada pela equação 5.14. No gráfico da Figura 5.14, apresentamos os valores calculados de W para os trios que se encaixavam nas condições impostas (inclusão

= nenhum *locus* onde $EV_m = 0$; e exclusão = dois ou mais *loci* onde $EV_m = 0$). O esperado para essa simulação são valores de W próximos de um para os casos de inclusão, e valores próximos de zero para os casos de exclusão.

Figura 5.14 - Dotplot de W dos dados artificiais. A probabilidade de paternidade baseada nas evidências de paternidade foi calculada para os casos simulados com 12 ou mais micro-haplótipos com qualidade e indícios de ser exclusão ou inclusão. Na legenda Simulação, *Exclusão* indica os trios onde o suposto pai foi simulado como o pai não verdadeiro e *Inclusão* indica os trios onde o suposto pai foi simulado como sendo o pai verdadeiro. O tamanho dos pontos é proporcional à quantidade de micro-haplótipos que foram utilizados nas análises. Observamos que todos os casos de *Inclusão* possuem valores de probabilidade próximos a um, e os casos de *Exclusão* foram determinados corretamente como exclusão, com exceção do T45_F e do T47_F.

Podemos observar no gráfico da Figura 5.14 que conforme o esperado, os casos de inclusão simulados possuem valores de W próximos de um. E em todos os casos, os valores de W foram superiores a 0,99. Com exceção dos trios T45_F e T47_F, os casos de exclusão simulados também indicaram exclusão, pois os valores de W estão próximos de zero.

Os casos de falsa inclusão (T45_F e T47_F) dados pelo nosso modelos podem ter ocorrido porque estamos utilizando um corte baixo para a quantidade mínima de micro-haplótipos analisados (12 ou mais), e conforme observado no Capítulo 4, o uso de poucos micro-haplótipos pode resultar em falsos positivos.

Testamos então a paternidade de todos os supostos pais contra todas as duplas de amostras mãe e plasma artificial. O esperado desta comparação eram casos de inclusão na diagonal, onde o suposto pai e a dupla mãe e plasma artificial pertencem ao mesmo grupo.

Esperamos também inclusões onde o suposto pai do trio criado para ser falso (trios seguidos por _F) está sendo comparado com o trio criado para ser verdadeiro, por exemplo: T34 e T34_F, o suposto pai T34_F deve indicar inclusão de paternidade com a dupla mãe e plasma artificial do T34. E em todas as outras comparações, o esperado são exclusões. O resultado dessa comparação está apresentado no gráfico da Figura 5.15.

Figura 5.15 - Heatmap da simulação suposto pai versus mãe e plasma artificial, mínimo de 12. Nessa simulação, testamos a paternidade do suposto pai versus a dupla mãe e plasma artificial. Nos casos onde o número de micro-haplótipos é menor que 12, definimos como *Menos de 12* e não analisamos porque a qualidade é baixa. Nos casos onde o número de micro-haplótipos é maior ou igual a 12, os dados são analisados. Nesses casos, se existir um micro-haplótipo onde $EV_m = 0$, é classificado como *Inconclusivo*. Se existir mais de um micro-haplótipo onde $EV_m = 0$, é classificado como *Exclusão*, e se não existir micro-haplótipo onde $EV_m = 0$, é classificado como *Inclusão*. Para os casos de *Inclusão*, a probabilidade de paternidade posterior (W) foi calculada, e esta é proporcional à transparência da cor.

Utilizamos como corte um mínimo de 12 micro-haplótipos com qualidade para serem analisados, e os casos abaixo desse corte, foram classificados como *Menos de 12*. Os casos onde existem mais de 12 micro-haplótipos, porém existem também dois ou mais micro-haplótipos indicando exclusão de paternidade ($EV_m = 0$), classificamos como *Exclusão*. Casos com mais de 12 micro-haplótipos, e apenas um micro-haplótipo indicando exclusão,

classificamos como *Inconclusivo*, pois não existem evidências muito fortes a respeito da exclusão. E por fim, casos com mais de 12 micro-haplótipos e nenhum micro-haplótipo indicando exclusão, classificamos como *Inclusão* e a probabilidade de paternidade (W) foi calculada. A transparência é proporcional ao valor de W , e quanto menos transparente, maior a probabilidade.

Figura 5.16 - Heatmap da simulação suposto pai versus mãe e plasma artificial, mínimo de 15. Nessa simulação, testamos a paternidade do suposto pai versus a dupla mãe e plasma artificial. Nos casos onde o número de micro-haplótipos é menor que 15, definimos como *Menos de 15* e não analisamos porque a qualidade é baixa. Nos casos onde o número de micro-haplótipos é maior ou igual a 15, os dados são analisados. Nesses casos, se existir um micro-haplótipo onde $EV_m = 0$, é classificado como *Inconclusivo*. Se existir mais de um micro-haplótipo onde $EV_m = 0$, é classificado como *Exclusão*, e se não existir micro-haplótipo onde $EV_m = 0$, é classificado como *Inclusão*. Para os casos de *Inclusão*, a probabilidade de paternidade posterior (W) foi calculada, e esta é proporcional à transparência da cor.

Observamos na Figura 5.15 que, conforme o esperado, a diagonal de inclusão está presente. Temos que o suposto pai está indicando inclusão com a mãe e plasma artificial correspondente. Além disso, o suposto pai do trio falso (trios seguidos de _F) também estão indicando inclusão com a mãe e plasma artificial dos seus correspondentes verdadeiros (ex: T34 e T34_F). O esperado para as comparações fora da diagonal eram casos de *Exclusão*,

Menos de 12 ou *Inconclusivos*. No entanto, ocorreram alguns casos de falsos positivos. Isso provavelmente ocorreu porque estamos utilizando nas análises apenas 12 micro-haplótipos. Como visto no Capítulo 4, quanto mais micro-haplótipos utilizados, menores as chances de ocorrerem falsos positivos.

No *heatmap* da Figura 5.16, repetimos o mesmo processo feito para criar o *heatmap* da Figura 5.15, com a diferença de que seriam utilizados 15 micro-haplótipos ou mais. Observamos que ao utilizar um mínimo de 15 micro-haplótipos, o número de falsos positivos diminui. Podemos observar, que com o aumento do número mínimo de micro-haplótipos sendo analisados, a ocorrência de falsos positivos diminui. Isto é mais um indicativo de que quanto maior a quantidade de micro-haplótipos utilizados nas análises, menores as chances de ocorrer uma falsa inclusão.

Conclusão

6.1 Considerações finais

Utilizando os dados do *1000 Genomes*, obtivemos todos os haplótipos possíveis encontrados nas 26 populações estudadas pelo banco de dados. Observamos das análises que realizamos, que todos os 20 micro-haplótipos escolhidos para o estudo são polimórficos, sendo que o menos polimórfico possui 6 haplótipos diferentes e o mais possui 48. Além disso, o maior intervalo entre a primeira base e a última base de um micro-haplótipo é de 140 pb, o que significa que a possibilidade de ocorrer um *crossing-over* dentro de um micro-haplótipo é extremamente pequena (taxas menores que 10^{-8}). Por essas razões, podemos concluir que essas regiões escolhidas possuem um grande potencial para serem utilizados como marcadores genéticos para testes de paternidade.

Realizamos também simulações de paternidade utilizando os dados do *1000 Genomes*, e podemos concluir com os estudos realizados, que quanto maior a quantidade de micro-haplótipos utilizados nas análises, maior a precisão do cálculo da probabilidade de paternidade. No entanto, se poucos micro-haplótipos forem utilizados, as chances de ocorrerem falsas inclusões de paternidade aumentam. Das nossas 52.000 simulações realizadas, podemos concluir que com uma quantidade de micro-haplótipos acima de 15, as chances de ocorrerem falsos positivos são menores do que 7 em 52.000.

Obtivemos os dados de sequenciamento dos trios de amostras suposto pai, mãe e plasma no banco de dados público da empresa Genomic Engenharia Molecular. Primeiramente, analisamos a qualidade dos dados de sequenciamento, e observamos que muitos dos trios possuíam poucos micro-haplótipos cobertos com qualidade para serem analisados. Optamos por utilizar um mínimo de 12 micro-haplótipos cobertos para seguirmos com as nossas análises, pois se utilizássemos 15, poucos trios teriam essa qualidade. Sabemos que existem

chances de ocorrerem falsos positivos com essa quantidade.

Para detectar os haplótipos que o feto herda apenas do suposto pai, utilizamos um intervalo de 1% a 12% de valores de frequências relativas. E para garantir a qualidade dos dados observados em um micro-haplótipo, excluimos este quando ocorrem erros de sequenciamento com frequências relativas superiores a 12%. Realizando este tipo de corte, muitos micro-haplótipos dos dados de sequenciamento dos trios, e principalmente do plasma, não passaram para as análises seguintes. Com este tipo de corte, diminuimos ainda mais a quantidade de micro-haplótipos cobertos nos trios. Mas garantimos a qualidade dos dados que estão sendo analisados.

Criamos as evidências de paternidade para obter indícios de que existe informação genética para indicar a inclusão ou a exclusão. Observamos nos nossos dados, que apenas um *locus* indicando a exclusão ($EV_m = 0$) não é evidência suficiente para que o suposto pai seja excluído como o pai verdadeiro, estes casos foram classificados como *Inconclusivos*. Por isso, determinamos que, para ser um caso de inclusão de paternidade, não pode ocorrer nenhum micro-haplótipo com $EV_m = 0$, e então calculamos a probabilidade de paternidade utilizando os dados populacionais. E para ser determinado como exclusão, é necessário que haja pelo menos dois micro-haplótipos onde $EV_m = 0$.

Realizamos o cálculo da probabilidade de paternidade baseado nas evidências de paternidade, e observamos que este método não foi capaz de distinguir os casos de exclusão dos casos de inclusão nos dados reais. Posteriormente, criamos um *dataset* a partir dos dados reais da Genomic, onde a fração fetal foi artificialmente adicionada, e para essas análises, utilizamos uma fração de 10%. Utilizando esse conjunto de dados, calculamos a probabilidade baseada nas evidências. Novamente, observamos que este método não foi capaz de distinguir os casos de exclusão dos casos de inclusão.

Utilizando o modelo clássico de cálculo da probabilidade, foi possível calcular a probabilidade de paternidade dos trios que possuíam 12 ou mais micro-haplótipos com qualidade, e que não fossem casos *Inconclusivos* (trios com um *locus* indicando exclusão). Nos dados reais da Genomic, observamos que a probabilidade de paternidade posterior (W) calculada, era superior a 99% nos casos de inclusão, o que é uma forte evidência de que o suposto pai é o pai verdadeiro. E não houveram casos de falsos positivos e nem falsos negativos. Ou seja, todos os casos de exclusão foram determinados corretamente, assim como os casos

de inclusão. Calculamos o índice de paternidade para os casos de mutação, e observamos que, embora possíveis, esses valores são da ordem de 5×10^{-5} ou menores.

Ainda utilizando os 60 trios de amostras da Genomic, realizamos uma simulação onde testamos todas as duplas mãe e plasma para paternidade contra todos os supostos pais. Observamos que ocorreram inclusões de paternidade inesperadas, e provavelmente, esses casos de falsos positivos ocorreram porque estamos utilizando 12 micro-haplótipos como o mínimo para as nossas análises. Ao utilizarmos o mínimo de 15 micro-haplótipos, a quantidade de falsos positivos diminuiu por não possuírem qualidade para serem analisados. Além disso, se mais micro-haplótipos possuísem qualidade para serem analisados, provavelmente ocorreria a detecção de uma ou mais regiões que indicariam a exclusão. Outra possibilidade é que os erros de sequenciamento, que estão dentro da faixa de detecção dos haplótipos fetais herdados do pai verdadeiro, formam haplótipos que são coincidentes com os haplótipos do suposto pai.

Calculamos W para os trios criados artificialmente a partir dos dados reais também, estes trios possuíam uma fração fetal artificial de 10%. Do cálculo realizado apenas para os trios que possuíam 12 ou mais micro-haplótipos cobertos e que não fossem casos inconclusivos, obtivemos dois casos de falsos positivos. Todos os outros trios foram determinados corretamente como inclusão (com valores de W superiores a 0.99) e exclusão (com valores de W próximos de zero). Provavelmente, os casos de falsa inclusão de paternidade aconteceram pela baixa qualidade do sequenciamento, que levou a uma grande quantidade de erros de sequenciamento e poucos micro-haplótipos cobertos por amostra. conforme observado no Capítulo 4, análises utilizando 15 micro-haplótipos ou mais, possuem chances menores de incluírem falsos positivos.

Ao testar a paternidade de todas as duplas mãe e plasma artificial contra supostos pais, observamos que os casos de inclusão estavam sendo determinados corretamente, quando havia 12 ou mais micro-haplótipos para serem analisados. No entanto, observamos também os casos de falsa inclusão. Ao utilizar um mínimo de 15 micro-haplótipos com qualidade para serem analisados, a quantidade de falsos positivos diminuiu. Mais uma vez, concluímos que estes casos estão ocorrendo devido à falta de mais micro-haplótipos cobertos com qualidade para serem utilizados nas análises, e também por conta dos erros de sequenciamento.

Realizamos uma simulação variando a fração fetal, para analisar se o *script* seria capaz de detectar frações fetais bem baixas. Podemos concluir que o método é capaz de detectar frações fetais entre 2% e 20%. Normalmente, a coleta de sangue da mãe é realizada após a 10^a semana de gravidez, e o esperado de fração fetal nesse estágio da gravidez é de aproximadamente 10%. Portanto, se houver a fração fetal, e a coleta de sangue for realizada no tempo certo de gravidez, a detecção dos haplótipos que o feto herda apenas do pai será possível.

Utilizando os trios de dados reais, que sabemos que são casos de inclusão de paternidade, observamos que, quando a fração fetal era detectada, ela não estava constante em todos os micro-haplótipos onde foi detectada. E em alguns casos em que surgiu um falso indício de exclusão onde só deveria ser observada a inclusão, a detecção não ocorreu porque a fração fetal estava abaixo do limite esperado (entre 1% e 12%). Provavelmente isso ocorreu por causa dos erros de sequenciamento que estavam atrapalhando as análises ou outros problemas relacionados à técnica de sequenciamento. Ao simular casos de inclusão de paternidade utilizando uma fração fetal de 10%, observamos que o método foi capaz de detectar os haplótipos fetais na frequência relativa (aproximadamente 5%) esperada para esse valor de fração fetal. E quando esta foi detectada, ela estava aproximadamente constante em todos os micro-haplótipos onde foi detectada.

Podemos concluir com as nossas análises, que os micro-haplótipos podem ser usados como marcadores genéticos para testes de paternidade. Idealmente, as análises devem compreender pelo menos 15 micro-haplótipos para que possamos obter valores altos de probabilidade de paternidade, e diminuir as chances de falsas inclusões. No entanto, quanto maior o número de micro-haplótipos, maiores as evidências para uma exclusão ou inclusão verdadeira, e menores as chances de falsas determinações.

6.2 Trabalhos futuros

Trabalhos futuros poderão ser realizados para descobrir novos micro-haplótipos para serem incluídos nas análises. Dessa forma, aumentamos a precisão dos cálculos e diminuimos as chances de ocorrerem falsos positivos.

Melhorias na etapa laboratorial para aumentar a cobertura dos micro-haplótipos, melhorar a qualidade das bases e diminuir a quantidade de erros de sequenciamento poderão ser feitas também. Dessa forma, as frações fetais serão melhor detectadas nos sequenciamentos do plasma e mais micro-haplótipos terão qualidade para serem utilizados nas análises.

Referências Bibliográficas

- [1] Lo YMD, Patel P, Wainscoat JS, Sampietro M, Gillmer MDG & Fleming KA. Prenatal sex determination by DNA amplification from maternal peripheral blood. *Lancet*. **1989**; *2(8676)*: 1363-1365.
- [2] Lo YMD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CWG & Wainscoat JS. Presence of fetal DNA in maternal plasma and serum. *Lancet*. **1997**; *350(9076)*: 485-487.
- [3] Ashoor G, Poon L, Syngelaki A, Mosimann B & Nicolaides KH. Fetal fraction in maternal plasma cell-free DNA at 11-13 weeks' gestation: effect of maternal and fetal factors. *Fetal Diagn Ther*. **2012**; *31(4)*: 237-243.
- [4] Wang E, Batey A, Struble C, Musci T, Song K & Oliphant A. Gestational age and maternal weight effects on fetal cell-free DNA in maternal plasma. *Prenat Diagn*. **2013**; *33(7)*: 662-666.
- [5] Rava RP, Srinivasan A, Sehnert AJ & Bianchi DW. Circulating fetal cell-free DNA fractions differ in autosomal aneuploidies and monosomy X. *Clin Chem*. **2014**; *60(1)*: 243-250.
- [6] Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, *et al*. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci Transl Med*. **2010**; *2(61)*: 3-16.
- [7] Palomaki GE, Kloza EM, Lambert-Messerlian GM, van der Boom D, Enrich M, Deciu C, *et al*. Circulating cell free DNA testing: are some test failures informative? *Prenat Diagn*. **2015**; *35(3)*: 289-293.

-
- [8] Ryan A, Baner J, Demko Z, Hill M, Sigurjonsson S, Baird ML & Rabinowitz M. Informatics-based, highly accurate, noninvasive prenatal paternity testing. *Genet Med*. **2013**; *15*(6): 473-477.
- [9] Saba L, Masala M, Capponi V, Marceddu G, Massida M & Rosatelli MC. Non-invasive prenatal diagnosis of beta-thalassemia by semiconductor sequencing: a feasibility study in the sardinian population. *Eur J Hum Genet*. **2017**; *25*(5): 600-607.
- [10] Straver R, Oudejans CBM, Sistermans EA & Reinders MJT. Calculating the fetal fraction noninvasive prenatal testing based on genome-wide nucleosome profiles. *Prenat Diagn*. **2016**; *36*(7): 614-621.
- [11] Akolekar R, Beta J, Ogilvie C & D'Antonio F. Procedure-related risk of miscarriage following amniocentesis and chorionic villus sampling: a systematic review and meta-analysis. *Ultrasound Obstet Gynecol*. **2015**; *45*(1): 16-26.
- [12] Thompson R, Zoppis S & McCord B. An overview of DNA typing methods for human identification: Past, present and future. *Methods Mol Biol*. **2012**; *830*: 3-16.
- [13] Dixon LA, Dobbins AE, Pulker HK, Butler JM, Vallone PM, Coble MD, *et al*. Analysis of artificially degraded DNA using STRs and SNPs - results of a collaborative European (EDNAP) exercise. *Forensic Sci Int*. **2006**; *164*(1): 33-44.
- [14] Tamaki K & Jeffreys AJ. Human tandem repeat sequences in forensic DNA typing. *Leg Med (Tokio)*. **2005**; *7*(4): 244-250.
- [15] Urquhart A, Kimpton CP, Downes TJ & Gill P. Variation in short tandem repeat sequences - a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Leg Med*. **1994**; *107*(1): 13-20.
- [16] Butler JM, Coble MD & Vallone PM. STRs vs. SNPs: Thoughts on the future of forensic DNA testing. *Forensic Sci Med Pathol*. **2007**; *3*(3): 200-205.
- [17] Tracey M. Short Tandem Repeat-based identification of individuals and parents. *Croat Med J*. **2001**; *42*(3): 233-238.
- [18] Grandell I, Samara R, & Tilmar AO. A SNP panel for identity and kinship testing using massive parallel sequencing. *Int J Legal Med*. **2016**; *130*(4): 905-914.

- [19] Debeljak M, Freed DN, Welch JA, Haley L, Beierl K, Iglehart BS, *et al.* Haplotype counting by Next-Generation Sequencing for ultrasensitive human DNA detection. *J Mol Diagn.* **2014**; *16*(5): 495-503.
- [20] Pakstis AJ, Fang R, Furtado MR, Kidd JR & Kidd KK. Mini-haplotypes as lineage informative SNPs and ancestry inference SNPs. *Eur J Hum Genet.* **2012**; *20*(11): 1148-1154.
- [21] Kidd KK, Pakstis AJ, Speed WC, Lagacé R, Chang J, Wootton S, *et al.* Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Sci Int Genet.* **2014**; *12*: 215-224.
- [22] Gibson G & Muse SV. A primer of genome science. *3. ed.* Sunderland, Estados Unidos da América: *Sinauer Associates*; **2009**.
- [23] Kidd KK, Pakstis AJ, Speed WC, Lagace R, Chang R, Wootton S & Ihuegbu N. Microhaplotype loci are a powerful new type of forensic marker. *Forensic Sci Int Gen Supplement Series.* **2013**; *4*(1): e123-e124.
- [24] Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SLB, Koroma NJ, *et al.* Developing a SNP panel for forensic identification of individuals. *Forensic Sci Int.* **2006**; *164*(1): 20-32.
- [25] Nachman MW & Crowell SL. Estimate of the mutation rate per nucleotide in humans. *Genetics.* **2000**; *156*(1): 297-304.
- [26] Guo F, Zhou Y, Song H, Zhao J, Shen H, Zhao B, *et al.* Next generation sequencing of SNPs using the HID-Ion AmpliSeqTM Identity Panel on the Ion Torrent PGMTM platform. *Forensic Sci Int Genet.* **2016**; *25*: 73-84.
- [27] Hollard C, Keyser C, Delabarde T, Gonzalez A, Lamego CV, Zvánigorosky V & Ludes B. Case report: on the use of HID-Ion AmpliSeqTM Ancestry Panel in a real forensic case. *Int J Legal Med.* **2017**; *131*(2): 351-358.
- [28] Richardson P. Special issue: next generation DNA sequencing. *Genes (Basel).* **2010**; *1*(3): 385-387.

- [29] Glenn TC. Field guide to next-generation sequencers. *Mol Ecol Resour.* **2011**; *11*(5): 759-769.
- [30] Loman NJ, Misra R, Dallman TJ, Constantinidou C, Gharbia SE, Wain J & Pallen MJ. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol.* **2012**; *30*(5): 434-439.
- [31] Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP & Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* **2012**; *13*: 341.
- [32] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* **2008**; *456*(7218): 53-59.
- [33] Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature.* **2011**; *475*(7356): 348-352.
- [34] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, *et al.* An integrated map of structural variation in 2504 human genomes. *Nature.* **2015**; *526*(7571): 75-81.
- [35] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, *et al.* A Global reference for human genetic variation. *Nature.* **2015**; *526*(7571): 68-74.
- [36] Varuzza L. Introdução à análise de dados de sequenciadores de nova geração. **2013**. [Disponível em: http://lvaruzza.com/files/apostila_bioinfo_2.1.pdf].
- [37] Li H, Handsaker B, Wysoker A, Fennel T, Ruan J & Homer N. The Sequence Alignment/Map format and Samtools. *Bioinformatics* **2009**, *25*: 2078-2079.
- [38] Nielsen R, Paul JS, Albrechtsen A & Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* **2011**; *12*(6): 443-451.
- [39] van Lint JH. A survey of perfect codes. *Rocky Mountain J Math.* **1975**; *5*(2): 199-224.

-
- [40] American Association of Blood Banks. Guidance for standard for relationship testing laboratories. 7. *ed.* Bethesda, Estados Unidos da América: *AABB*; **2006**.

Apêndice A

Populações do *1000 Genomes*

Os dados do *1000 Genomes* foram obtidos de 26 populações diferentes, divididas em cinco super-populações: AFR, AMR, EAS, EUR e SAS. Na terceira fase do projeto, o banco de dados obteve informações de variantes genéticas de 2504 indivíduos.

Tabela A.1 - Lista das super-populações e populações de onde as amostras do *1000 Genomes* provêm, com suas respectivas siglas e quantidade de indivíduos.

Super-pop	Sigla População	População	Nº de amostras
AFR	ACB	<i>African Caribbean in Barbados</i>	96
AFR	ASW	<i>African Ancestry in Southwest US</i>	61
AFR	ESN	<i>Esan in Nigeria</i>	99
AFR	GWD	<i>Gambian in Western Division, The Gambia</i>	113
AFR	LWK	<i>Lugya in Webuye, Kenya</i>	99
AFR	MSL	<i>Mende in Sierra Leone</i>	85
AFR	YRI	<i>Yoruba in Abadan, Nigeri</i>	108
AMR	CLM	<i>Colombian in Medellin, Colombia</i>	94
AMR	MXL	<i>Mexican Ancestry in Los Angeles, California</i>	64
AMR	PEL	<i>Peruvian in Lima, Peru</i>	85
AMR	PUR	<i>Puerto Rican in Puerto Rico</i>	104
EAS	CDX	<i>Chinese Dan in Xishuangbanna, China</i>	93
EAS	CHB	<i>Han Chinese in Beijing, China</i>	103
EAS	CHS	<i>Han Chinese South</i>	105
EAS	JPT	<i>Japanese in Tokyo, Japan</i>	104
EAS	KHV	<i>Kinh in Ho Chi Minh City, Vietnam</i>	99
EUR	CEU	<i>Utah residents with Northern and Western European ancestry</i>	99
EUR	FIN	<i>Finnish in Finland</i>	99
EUR	GBR	<i>British in England and Scotland</i>	91
EUR	IBS	<i>Iberian population in Spain</i>	107
EUR	TSI	<i>Toscani in Italy</i>	107
SAS	BEB	<i>Bangali in Bangladesh</i>	86
SAS	GIH	<i>Gujarati Indian in Houston, TX</i>	103
SAS	ITU	<i>Indian Telugu in the UK</i>	102

Continua na próxima página. . .

Tabela A.1 - Continuação

Super-pop	Sigla População	População	Nº de amostras
SAS	PJL	<i>Punjabi in Lahore, Pakistan</i>	96
SAS	STU	<i>Sri Lankan Tamil in the UK</i>	102

Apêndice B

Haplótipos do micro-haplótipo

Seguindo o mesmo método utilizado para obter os haplótipos descrito no Capítulo 2.2, obtivemos os haplótipos e a frequência dos mesmos para os outros 19 micro-haplótipos analisados.

Tabela B.1 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M02 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M02H01	AGCACGT	1	0	0	0	0	1
M02H02	AGCGCGT	1679	564	212	216	378	309
M02H03	AGTAAAC	1	0	0	0	0	1
M02H04	GACAAGT	1	0	0	1	0	0
M02H05	GACGCGT	1084	292	165	85	210	332
M02H06	GATAAAC	2138	414	314	665	416	329
M02H07	GATGAAC	99	52	1	40	0	6
M02H08	GATGCAC	1	0	1	0	0	0
M02H09	GATGCGT	4	0	1	1	2	0

Tabela B.2 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M03 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M03H01	AAGATAACGCG	1	1	0	0	0	0

Continua na próxima página...

Tabela B.2 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M03H02	AAGTCAACGCG	1	0	0	0	1	0
M03H03	AAGTCAATAAA	3	1	1	0	1	0
M03H04	AAGTCAATAAG	1	1	0	0	0	0
M03H05	AAGTCAATGCG	1	0	0	0	1	0
M03H06	AAGTTAACGCG	14	3	1	4	4	2
M03H07	AAGTTAATAAA	1	1	0	0	0	0
M03H08	AAGTTAATAAG	5	0	0	1	1	3
M03H09	ACGTTAGTAAG	646	309	72	8	174	83
M03H10	AGGATAACGCG	2	1	1	0	0	0
M03H11	AGGTCAGTAAA	977	361	130	13	248	225
M03H12	AGGTTAGCGCG	1	1	0	0	0	0
M03H13	AGGTTAGTAAA	1441	297	264	406	197	277
M03H14	AGGTTAGTAAG	261	43	23	169	7	19
M03H15	AGGTTGGTAAG	377	92	63	132	57	33
M03H16	GAAATAACGCG	2	0	1	0	1	0
M03H17	GAGATAACAAA	1	0	0	0	0	1
M03H18	GAGATAACGCG	1257	207	138	273	311	328
M03H19	GAGATAATAAA	1	0	0	0	0	1
M03H20	GAGATAGCGCG	1	0	0	0	0	1
M03H21	GAGATAGTAAG	1	0	0	1	0	0
M03H22	GAGTCAATGCG	1	0	0	0	0	1
M03H23	GAGTTAACGCG	1	1	0	0	0	0
M03H24	GAGTTAATAAA	2	2	0	0	0	0
M03H25	GAGTTAATAAG	2	1	0	0	0	1
M03H26	GCGATAACGCG	3	0	0	0	3	0
M03H27	GCGTTAATAAG	1	0	0	0	0	1
M03H28	GGGATAACGCG	1	0	0	0	0	1
M03H29	GGGATAGCGCG	2	0	0	1	0	1

Tabela B.3 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M04 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M04H01	CAACACAT	2	2	0	0	0	0
M04H02	CAACACTC	25	23	1	0	1	0
M04H03	CAACACTT	993	166	159	140	234	294
M04H04	CAACGCTT	261	4	50	79	59	69
M04H05	CAATACAT	238	14	19	97	64	44

Continua na próxima página...

Tabela B.3 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M04H06	CAATACTT	14	1	0	2	7	4
M04H07	CCACACTC	264	61	76	5	113	9
M04H08	CCACACTT	1	0	1	0	0	0
M04H09	TAACACTC	20	1	2	17	0	0
M04H10	TAACACTT	973	276	77	272	139	209
M04H11	TAATACAT	44	39	1	2	2	0
M04H12	TAATACTC	2	1	0	0	1	0
M04H13	TAATACTT	614	174	87	219	83	51
M04H14	TAGCACTT	133	39	2	58	3	31
M04H15	TAGCATTT	366	109	46	5	132	74
M04H16	TCATACAT	764	390	136	13	144	81
M04H17	TCATACTC	293	22	37	99	24	111
M04H18	TCATACTT	1	0	0	0	0	1

Tabela B.4 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M05 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M05H01	ACGCGT	221	208	9	0	4	0
M05H02	GCAAGA	1	0	0	1	0	0
M05H03	GCAAGT	1803	342	293	384	454	330
M05H04	GCGAGT	683	216	83	62	148	174
M05H05	GCGCAA	1	0	0	1	0	0
M05H06	GCGCGT	471	225	48	5	79	114
M05H07	GGAAGT	1	0	0	0	1	0
M05H08	GGGCAA	1827	331	261	555	320	360

Tabela B.5 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M06 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M06H01	CGAATAT	24	23	1	0	0	0
M06H02	CGAGGAT	1898	320	310	302	580	386
M06H03	CGAGGTC	2	0	0	0	0	2
M06H04	TGAGGAT	1	0	0	1	0	0
M06H05	TGGGGTC	2021	687	251	387	282	414
M06H06	TTGGGTC	1062	292	132	318	144	176

Tabela B.6 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M07 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M07H01	ATCAGGGGCGCG	43	43	0	0	0	0
M07H02	ATCTAGGGCGCG	1	1	0	0	0	0
M07H03	ATCTGAGGCGCG	1	0	0	0	0	1
M07H04	ATCTGGGATATC	5	0	2	2	1	0
M07H05	ATCTGGGGCACG	1	0	0	0	0	1
M07H06	ATCTGGGGCGCG	2790	519	370	630	587	684
M07H07	GCGTAAAATATC	417	2	14	249	32	120
M07H08	GCGTAAAGCGCG	3	0	0	1	0	2
M07H09	GCGTAAGATATC	1501	532	293	126	382	168
M07H10	GCGTAAGGCGCG	12	3	3	0	4	2
M07H11	GCGTGAGATATC	3	3	0	0	0	0
M07H12	GTGTAGGACATC	231	219	12	0	0	0

Tabela B.7 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M08 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M08H01	CGACGGTATA	1	0	0	1	0	0
M08H02	CGACGGTATG	1296	189	247	123	374	363
M08H03	CGGAAACATG	1260	213	217	354	244	232
M08H04	CGGAAACGTG	165	159	6	0	0	0
M08H05	CGGAGGTATG	6	0	0	6	0	0
M08H06	TAGCAACATG	1	0	0	1	0	0
M08H07	TAGCGGTATG	103	2	1	53	1	46
M08H08	TGACAACATA	2	0	1	1	0	0
M08H09	TGACAACATG	11	0	2	0	9	0
M08H10	TGACGGTACA	42	40	2	0	0	0
M08H11	TGACGGTACG	230	209	16	1	2	2
M08H12	TGACGGTATA	1177	137	134	376	292	238
M08H13	TGACGGTATG	140	27	5	38	2	68
M08H14	TGGAAATATG	48	45	3	0	0	0
M08H15	TGGAGGTATG	1	0	0	1	0	0
M08H16	TGGCAATATG	524	301	60	52	82	29
M08H17	TGGCGGTATG	1	0	0	1	0	0

Tabela B.8 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M09 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M09H01	GCCGCAT	8	5	1	0	2	0
M09H02	GCCGCGC	749	406	90	102	72	79
M09H03	GCCGCGT	7	0	2	0	4	1
M09H04	GCCGTGC	1006	196	150	229	127	304
M09H05	GCCTCGC	6	4	0	1	0	1
M09H06	GCCTTGC	1261	198	185	220	410	248
M09H07	GCTGCAT	15	6	1	4	2	2
M09H08	GGTGCAT	7	1	1	2	3	0
M09H09	GGTGCGC	2	0	0	0	2	0
M09H10	TGTGCAT	1610	336	224	450	277	323
M09H11	TGTGCGC	252	168	27	0	44	13
M09H12	TGTGCGT	85	2	13	0	63	7

Tabela B.9 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M10 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M10H01	ACCCCGACC	5	0	1	2	1	1
M10H02	ACCCCGGCC	10	1	0	3	2	4
M10H03	ACCCCGGCT	10	2	3	2	3	0
M10H04	ACCCTGGCT	1	1	0	0	0	0
M10H05	AGTCCTACC	1	0	0	0	0	1
M10H06	AGTTCGGCT	4	1	0	0	1	2
M10H07	AGTTCTACC	2517	522	340	737	488	430
M10H08	GCCCCGGCT	2	2	0	0	0	0
M10H09	GCCCCGGGT	2364	712	341	264	509	538
M10H10	GCCCTGGGT	88	79	9	0	0	0
M10H11	GGTTCGGGT	1	0	0	0	1	0
M10H12	GGTTCTGCT	1	0	0	0	0	1
M10H13	GGTTCTGGT	4	2	0	0	1	1

Tabela B.10 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M11 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M11H01	ACAACACA	2	2	0	0	0	0
M11H02	ACAACACG	3	0	0	3	0	0
M11H03	ACAATACA	9	1	1	0	7	0
M11H04	ACAATGCA	1	0	0	0	1	0
M11H05	ACAATGCG	76	1	11	1	42	21
M11H06	ACGACACA	4	0	0	1	3	0
M11H07	ACGACACG	7	0	1	0	5	1
M11H08	ACGATACA	40	0	5	4	12	19
M11H09	ACGATACG	1	0	0	0	0	1
M11H10	ACGGCAAA	4	4	0	0	0	0
M11H11	ACGGTGCA	2	2	0	0	0	0
M11H12	ATAACAAA	2	2	0	0	0	0
M11H13	ATAACACA	152	64	19	35	27	7
M11H14	ATAACACG	1418	376	282	310	216	234

Continua na próxima página. . .

Tabela B.10 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M11H15	ATAACGCA	1	1	0	0	0	0
M11H16	ATAACGCG	20	1	6	12	0	1
M11H17	ATAATACA	1242	266	166	234	231	345
M11H18	ATAATACG	5	0	1	0	2	2
M11H19	ATAATGCA	146	137	3	1	2	3
M11H20	ATAATGCG	507	50	70	54	155	178
M11H21	ATAGTACA	1	1	0	0	0	0
M11H22	ATAGTGCA	9	9	0	0	0	0
M11H23	ATGACACA	2	0	0	1	1	0
M11H24	ATGACACG	29	2	2	12	4	9
M11H25	ATGATACA	9	2	0	5	0	2
M11H26	ATGATGCG	2	0	0	0	2	0
M11H27	ATGGCACA	1	1	0	0	0	0
M11H28	ATGGCACG	10	10	0	0	0	0
M11H29	ATGGTGCA	1	1	0	0	0	0
M11H30	CCAACAAA	94	88	5	0	1	0
M11H31	CCAACACA	8	4	1	1	2	0
M11H32	CCAACACG	13	6	0	7	0	0
M11H33	CCAATACA	25	15	1	0	8	1
M11H34	CCAATGCA	21	21	0	0	0	0
M11H35	CCAATGCG	7	0	3	0	3	1
M11H36	CCAGTGCA	8	8	0	0	0	0
M11H37	CCGACACA	809	75	105	264	222	143
M11H38	CCGACACG	55	3	7	17	23	5
M11H39	CCGATACA	38	25	0	6	6	1
M11H40	CCGATGCG	69	2	0	40	24	3
M11H41	CCGGCAAA	39	38	1	0	0	0
M11H42	CCGGCACA	57	56	1	0	0	0
M11H43	CCGGTGCA	5	5	0	0	0	0
M11H44	CTAACACA	1	1	0	0	0	0
M11H45	CTAACACG	37	29	3	0	5	0
M11H46	CTAATACA	3	0	0	0	2	1
M11H47	CTAATGCA	12	12	0	0	0	0
M11H48	CTAATGCG	1	1	0	0	0	0

Tabela B.11 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M12 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M12H01	CATCAA	1	1	0	0	0	0
M12H02	CATCGG	1513	297	222	324	385	285
M12H03	CATGGA	2	0	0	1	0	1
M12H04	CGTCGG	6	1	4	0	1	0
M12H05	TAGGGA	2	1	0	0	0	1
M12H06	TATCGA	115	26	6	31	4	48
M12H07	TATCGG	358	104	67	143	4	40
M12H08	TATGAA	1	1	0	0	0	0
M12H09	TATGGA	1866	380	300	364	408	414
M12H10	TATGGG	4	0	0	0	4	0
M12H11	TGGCAA	409	212	29	14	67	87
M12H12	TGGCGA	3	2	1	0	0	0
M12H13	TGTCAA	721	291	65	130	133	102
M12H14	TGTCTGA	4	4	0	0	0	0
M12H15	TGTCTGG	1	0	0	1	0	0
M12H16	TGTGGA	2	2	0	0	0	0

Tabela B.12 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M13 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M13H01	ACGCCACGC	2	0	0	0	0	2
M13H02	ACGTCACGC	22	0	1	3	0	18
M13H03	ACGTCACGT	1098	301	175	117	300	205
M13H04	ACGTTGTGC	3	0	1	0	1	1
M13H05	ACTCCACAC	4	3	1	0	0	0
M13H06	ACTCCACGC	2447	490	339	520	573	525
M13H07	ACTCTGTGC	1	0	1	0	0	0
M13H08	ACTTTGTGC	1	0	0	0	1	0
M13H09	ATGTTGTGC	17	0	0	14	2	1
M13H10	GCTCCACGC	1	1	0	0	0	0
M13H11	GTGTTGTGC	1412	527	176	354	129	226

Tabela B.13 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M14 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M14H01	GGGGCGGCAATAG	108	101	6	0	1	0
M14H02	GGGGCTACCATAG	245	31	41	1	142	30
M14H03	GGGGCTATAACAA	1	0	1	0	0	0
M14H04	GGGGCTGCAACAG	2	1	0	1	0	0
M14H05	GGGGCTGCAATAA	1	0	0	1	0	0
M14H06	GGGGCTGCAATAG	2593	411	367	720	591	504
M14H07	GGGGCTGCCATAG	1	0	0	0	0	1
M14H08	GGTGCTGCAATAG	2	2	0	0	0	0
M14H09	TAGACTATAACAA	2	2	0	0	0	0
M14H10	TATACTACAACAA	1	1	0	0	0	0
M14H11	TATACTATAACAA	1433	689	210	96	197	241
M14H12	TATACTATAACCA	453	16	63	106	71	197
M14H13	TATACTATACCAA	63	59	4	0	0	0
M14H14	TATACTGTAACAA	2	2	0	0	0	0
M14H15	TATAGTATAACAA	93	0	2	83	3	5
M14H16	TATGCTATAACAA	7	7	0	0	0	0
M14H17	TATGCTATAACCA	1	0	0	0	1	0

Tabela B.14 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M15 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M15H01	ACCCCGA	2	0	0	2	0	0
M15H02	ACCCCGG	610	14	69	287	143	97
M15H03	ACCCGCA	487	185	68	182	3	49
M15H04	ACCCGCG	5	0	1	2	2	0
M15H05	ACCTCGG	26	22	1	2	0	1
M15H06	ACCTGCA	20	11	1	0	1	7
M15H07	ACTCCCA	5	3	0	2	0	0
M15H08	ACTCCGG	2	0	1	0	0	1
M15H09	ACTCGCA	1320	405	186	108	442	179
M15H10	ACTTCGG	3	0	1	0	0	2

Continua na próxima página...

Tabela B.14 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M15H11	ATCCCGG	512	84	79	129	77	143
M15H12	ATCCGCA	2	1	0	0	1	0
M15H13	ATCTCGA	1	1	0	0	0	0
M15H14	ATCTCGG	440	165	29	69	37	140
M15H15	ATCTGCA	136	16	24	2	41	53
M15H16	ATCTGCG	1	0	1	0	0	0
M15H17	ATTGCA	10	0	0	10	0	0
M15H18	GCCCCGG	3	0	1	0	2	0
M15H19	GCCCGCA	75	0	8	0	54	13
M15H20	GCCTCGG	2	2	0	0	0	0
M15H21	GCCTGCA	36	22	4	2	3	5
M15H22	GCTCCCA	1	1	0	0	0	0
M15H23	GCTCCGA	90	86	4	0	0	0
M15H24	GCTCCGG	1	0	0	0	1	0
M15H25	GCTCGCA	916	273	132	58	188	265
M15H26	GCTTCGG	7	7	0	0	0	0
M15H27	GCTTGCA	57	1	29	27	0	0
M15H28	GTCCCGG	3	0	1	2	0	0
M15H29	GTCCGCA	1	0	0	1	0	0
M15H30	GTCTCGG	19	10	1	0	6	2
M15H31	GTCTGCA	214	13	53	122	5	21
M15H32	GTTCGCA	1	0	0	1	0	0

Tabela B.15 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M16 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M16H01	AGGCGCG	836	279	126	270	36	125
M16H02	AGGCGTG	2	2	0	0	0	0
M16H03	CCATTCG	1568	390	178	192	496	312
M16H04	CCGCGCG	1	0	0	1	0	0
M16H05	CCGTTTCG	1	0	0	1	0	0
M16H06	CGGCGCC	137	3	24	45	14	51
M16H07	CGGCGCG	1973	513	281	499	258	422
M16H08	CGGCGTC	8	1	1	0	5	1
M16H09	CGGCGTG	482	134	84	0	197	67

Tabela B.16 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M17 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M17H01	CCAC	268	123	18	29	55	43
M17H02	CCAT	2	1	0	0	1	0
M17H03	CCGC	189	128	13	2	32	14
M17H04	CCGT	1565	305	316	408	320	216
M17H05	CTAC	2	1	0	0	0	1
M17H06	CTGC	570	48	50	68	109	295
M17H07	CTGT	12	11 0	0	1	0	
M17H08	TCAC	1351	216	204	306	368	257
M17H09	TCGC	250	216	16	2	8	8
M17H10	TCGT	523	58	65	188	88	124
M17H11	TTAC	6	2	0	3	0	1
M17H12	TTGC	233	183	11	1	21	17
M17H13	TTGT	37	30 1	1	3	2	

Tabela B.17 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M18 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M18H01	CATCATCTG	3	3	0	0	0	0
M18H02	CATCCTCCG	1	1	0	0	0	0
M18H03	CATCCTCTG	3	3	0	0	0	0
M18H04	CGACACGCA	1	0	1	0	0	0
M18H05	CGACACGCG	1	1	0	0	0	0
M18H06	CGACATCCA	1	0	0	0	0	1
M18H07	CGACCCCTG	44	41	3	0	0	0
M18H08	CGACCCGCG	2126	541	396	213	598	378
M18H09	CGACCTGCG	80	77	2	0	1	0
M18H10	CGACTCGCG	2	0	2	0	0	0
M18H11	TATCACCCA	1	0	1	0	0	0
M18H12	TATCATCCA	1969	551	247	465	313	393
M18H13	TATCATCCG	7	2	0	5	0	0
M18H14	TATCCTCCG	1	1	0	0	0	0

Continua na próxima página...

Tabela B.17 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M18H15	TATTATCTG	694	32	37	325	94	206
M18H16	TGTCATCCA	74	69	5	0	0	0

Tabela B.18 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M19 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M19H01	AGTAAT	3	1	1	0	1	0
M19H02	AGTTAT	88	79	5	0	4	0
M19H03	AGTTGC	2775	300	389	766	625	695
M19H04	AGTTGT	130	117	9	3	1	0
M19H05	CACAAC	2	0	1	0	0	1
M19H06	CACAAT	2009	825	289	239	375	281
M19H07	CACAGC	1	0	0	0	0	1

Tabela B.19 - Lista dos haplótipos encontrados nos banco *1000 Genomes* para o micro-haplótipo M20 e a frequência deles levando em conta todas as populações e as super-populações .

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M20H01	CCAACCCCTACT	2	0	0	0	2	0
M20H02	CCAACCCCTCAC	5	1	1	0	0	3
M20H03	CCGACCCCTCAC	1	0	0	0	0	1
M20H04	CCGACCCTAACT	1	1	0	0	0	0
M20H05	CCGCCCCCTCAC	4	0	0	1	0	3
M20H06	CCGCTCCCTCAC	5	1	1	0	0	3
M20H07	CCGCTCTCTCAC	1	0	0	0	0	1
M20H08	CCGCTCTTAACT	2567	707	367	349	626	518
M20H09	CCGCTGTTAACT	17	16	1	0	0	0
M20H10	CTAACCCCTACT	1	1	0	0	0	0
M20H11	CTAACCCCTCAC	37	5	5	12	6	9
M20H12	CTAACCTTAACT	4	1	1	2	0	0

Continua na próxima página. . .

Tabela B.19 - Continuação

ID haplótipo	Haplótipo	TODOS	AFR	AMR	EAS	EUR	SAS
M20H13	CTGCCCCCTCAC	5	0	0	3	1	1
M20H14	CTGCTCCCTCAC	5	0	1	0	3	1
M20H15	CTGCTCTTAAC	27	3	5	10	8	1
M20H16	TCAACCCCTCAC	1	1	0	0	0	0
M20H17	TCGACCCCTCAC	1	0	0	0	0	1
M20H18	TCGCTCCCTCAC	2	2	0	0	0	0
M20H19	TCGCTCTCTCAC	1	0	0	0	0	1
M20H20	TCGCTCTTAAC	14	6	0	0	1	7
M20H21	TCGCTCTTTCAC	1	0	0	0	0	1
M20H22	TTAACCCCTACC	2	0	1	0	0	1
M20H23	TTAACCCCTACT	25	4	5	7	6	3
M20H24	TTAACCCCTCAC	1780	311	265	565	273	366
M20H25	TTAACCCCTCAT	1	0	0	1	0	0
M20H26	TTAACCCCTAAC	2	0	0	1	0	1
M20H27	TTAACCCCTAAC	19	7	1	3	4	4
M20H28	TTAACCTTAACC	1	0	1	0	0	0
M20H29	TTAACCTTAAC	63	18	6	17	12	10
M20H30	TTAATCTTAAC	11	2	2	1	6	0
M20H31	TTAATCTTACAC	2	1	0	0	1	0
M20H32	TTACCCCTCAC	1	0	0	0	0	1
M20H33	TTACTCCCTCAC	222	200	9	0	2	11
M20H34	TTACTCTTAAC	2	0	0	1	1	0
M20H35	TTGACCCCTACT	2	0	0	0	2	0
M20H36	TTGACCCCTCAC	7	0	0	2	5	0
M20H37	TTGCCCCCTCAC	3	0	1	2	0	0
M20H38	TTGCTCCCTCAC	40	4	3	8	10	15
M20H39	TTGCTCTCTCAC	9	2	1	2	3	1
M20H40	TTGCTCTTAAC	108	25	17	18	34	14
M20H41	TTGCTCTTACAC	6	3	0	3	0	0

Apêndice C

Índice de Paternidade

Temos que A, B, C e D representam os haplótipos e a e b são as frequências dos haplótipos A e B respectivamente.

Tabela C.1 - Lista dos valores de IP_m para diferentes combinações de haplótipos do suposto pai, mãe e filho

Filho	Mãe	Suposto Pai	IP_m
AA	AA	AA	$\frac{1}{a}$
		AB	$\frac{1}{2a}$
	AB	AA	$\frac{1}{a}$
		AB	$\frac{1}{2a}$
		AC	$\frac{1}{2a}$
AB	AB	AA	$\frac{1}{a+b}$
		AB	$\frac{1}{a+b}$
		AC	$\frac{1}{2(a+b)}$
	BB	AA	$\frac{1}{a}$
		AB	$\frac{1}{2a}$
		AC	$\frac{1}{2a}$
	BC	AA	$\frac{1}{a}$
		AB	$\frac{1}{2a}$
		AC	$\frac{1}{2a}$
	BD	AC	$\frac{1}{2a}$