Marcel Rodrigues Ferreira

Epigenética e fenotipagem forense por DNA: aplicação da metilação de ilhas CpG e promotores na predição de idade e pigmentação da pele a partir de tecido epitelial

Relatório de Pós-doutorado realizado na Universidade Estadual Paulista (UNESP), Faculdade de Medicina de Botucatu.

Supervisor(a): Prof. Dr. Erick da Cruz Castelli

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - 88887.848127/2023-00

Botucatu-SP

2025

Sumário

Sumário i

Lista de figuras iii

Lista de tabelas vi

RESUMO vii

1. INTRODUÇÃO 1

1.1 Genética forense 1

1.1 Fenotipagem forense de DNA 2

1.2 Epigenética 4

1.3 Sequenciamento do genoma completo via Oxford Nanopore Technologies 6

1.4 Justificativa para este projeto 8

2. OBJETIVOS 9

3. MATERIAIS E MÉTODOS 9

3.1 Parte 1 – Coleta de amostras e determinação de métodos de extração de DNA 9

3.1.1 Coleta de amostras e caracterização fenotípica da amostra populacional 9

3.1.2 Extração de DNA 10

3.2 Análises computacionais 11

3.2.1 Dados públicos 11

3.2.2 Amostras utilizadas 12

3.2.3 Coordendas genômicas 14

3.2.4 Chamada de base, alinhamento e chamada de variantes 17

3.2.5 Determinação de haplótipos longos 17

3.2.6 Análise de metilação 17

4. RESULTADOS E DISCUSSÃO 18

4.1 Amostras coletadas 18

4.1.1 Características fenotípicas das amostras 18

4.1.2 Quantificação das amostras 24

4.2 Estabelecimento dos procedimentos computacionais com a amostra HG002 26

4.2.1 Pipeline para análise de dados de sequenciamento via Oxford Nanopore Technologies 26

4.2.2 Pipeline para chamada de variantes....29

4.2.3 Procedimentos para análise da metilação....34

4.3 Análises bioinformáticas em dados públicos do Projeto 1000 Genomes....39

4.3.1 Análise da cobertura dos genes....40

4.3.2 Haplótipos longos dos genes de pigmentação....43

4.3.3 Metilação de genes de pigmentação....53

4.3.4 Metilação diferencial entre haplótipos....54

4.3.5 Entropia de metilação....57

5 CONCLUSÃO....60

REFERÊNCIAS....60

ANEXO....67

ORIENTAÇÕES/COORIENTAÇÕES....67

DISCIPLINAS MINISTRADAS COM A RESPECTIVA CARGA HORÁRIA E/OU REALIZAÇÃO DE ATIVIDADES DE EXTENSÃO, COMENTÁRIOS DO SUPERVISOR)....67

PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS NO PERÍODO....67

LISTA DE PUBLICAÇÕES NO PERÍODO....69

Lista de figuras

Figura 1 - Principais benefícios proporcionados pelo sequenciamento de leitura longa para a genética forense. Figura republicada de (Ferreira et al., 2025). 2

Figura 2 - Modificações de citosina antes e depois da conversão de bissulfito. Figura retirada de (Ferreira et al., 2025). 5

Figura 3 - Visão geral do sequenciamento Oxford Nanopore. Os elementos gráficos foram obtidos em https://smart.servier.com/, enquanto as imagens do equipamento Oxford Nanopore foram baixadas de https://nanoporetech.com/. 7

Figura 4 - Gráfico de barras dos tipos de pele coletados. 18

Figura 5 - Gráfico de barras da cor dos olhos das amostras coletadas.....19

Figura 6 - Gráfico de barras das cores de cabelo das amostras coletadas.....19

Figura 7 - Gráfico de barras dos tipos de cabelo das amostras coletadas. 20

Figura 8 - Gráfico de barras dos tipos de cabelo das amostras coletadas. 20

Figura 9 - Gráfico de barras da porcentagem de grisalhos nas amostras coletadas.....21

Figura 10 - Gráfico de barras da classificação do IBGE das amostras coletadas.....21

Figura 11 - Gráfico de barras da classificação do IBGE dos pais das amostras coletadas. 22

Figura 12 - Gráfico de barras das origens dos avós das amostras coletadas. 22

Figura 13 – Gráfico violino da estatura (cm), idade (ano) e peso (kg) das amostras coletadas. 23

Figura 14 - Gráfico de barras do grau de escolaridade das amostras coletadas. 23

Figura 15 -Gráfico de barras da média salarial das amostras coletadas.....24

Figura 16 - Fluxograma de procedimentos para análise de dados de sequenciamento Oxford Nanopore. 28

Figura 17 - Fluxograma de procedimentos para chamada de variantes em dados de sequenciamento Oxford Nanopore via freebayes e Clair3. 30

Figura 18 - Visualização da cobertura por posição dos genes para a amostra HG002 do GIAB.....31

Figura 19 - Visualização da região do gene TYR no IGV. As trilhas superiores contêm as variantes para Freebayes e Clair3. 32

Figura 20 - Visualização da região do gene MC1R no IGV para a comparação dos VCFs com o fornecido por GIAB.....33

Figura 21 – Visualização da região do gene TYR no IGV para a comparação dos VCFs com o fornecido por GIAB.....33

Figura 22 - Visualização da região do gene OCA2 no IGV para a comparação dos VCFs com o fornecido por GIAB.....33

Figura 23 - Visualização da região do gene HERC2 no IGV para a comparação dos VCFs com o fornecido por GIAB.....34

Figura 24 - Visualização da região do gene MYEF2 no IGV para a comparação dos VCFs com o fornecido por GIAB.....34

Figura 25 - bedMethyl gerado por modkit utilizando o resultado de chamada de bases modificadas da amostra HG002 do GIAB. Nomes de colunas foram inseridos manualmente. As colunas marcadas em vermelho foram expandidas do seu formato original para melhor visualização. A coluna C (start position) é a posição da base modificada, devido ao formato de posição iniciando em 0.....36

Figura 26 - Fluxograma de procedimentos para análise de modificações de bases em dados de sequenciamento Oxford Nanopore. É necessário utilizar uma configuração adequada para a detecção das modificações de interesse (Ex: 5mC ou 5hmC) já no processo de chamada de base. 37

Figura 27 - Arquivo BAM com as modificações visualizado no IGV para a região do gene MC1R utilizando a amostra HG002 do GIAB. Foram selecionadas as opções de visualização color aligment by base modification e "*squished*".....38

Figura 28 - Arquivo BAM com as modificações visualizado no IGV para a região do gene MC1R (chr16:89,919,709-89,919,759) utilizando a amostra HG002 do GIAB. Paineis de visualização do bedMethyl, cobertura do BAM (ambas as fitas) e duas bases diferentes (uma de cada fita).....39

Figura 29 - Cobertura média por gene em cada amostra. Amostras foram ordenadas da maior para menor cobertura.....41

Figura 30 - Distribuição de valores de cobertura média por amostra. Amostras foram ordenadas da maior para menor cobertura.....43

Figura 31 - Número de haplótipos encontrados por gene/amostra. 45

Figura 32 - Número de contigs encontrados para cada haplótipo dos genes nas 57 amostras analisadas.....46

Figura 33 - Porcentagem coberta do gene pelos haplótipos (considerando todos os contigs) encontrados. 48

Figura 34 - Cobertura dos haplótipos de ASIP para as diferentes amostras analisadas. Abaixo contem um inserto representando os exóns do gene. 50

Figura 35 - Cobertura dos haplótipos de DEF8 para as diferentes amostras analisadas 52

Figura 36 - Perfil de metilação do gene SLC24A4 na amostra GM19238. 54

Figura 37 - Regiões diferencialmente metiladas (DMR) comparando-se os dois haplótipos de cada indivíduo/gene. Contagens apresentadas com escala log10. 55

Figura 38 - Screenshot do IGV na região chr8:3,744,135-3,744,216 do gene CSMD1. A região contém exemplo de 3 diferentes citosinas com metilação diferencial entre os haplótipos do indivíduo HG02885.....56

Figura 39 - Metilação ao longo do gene CSMD1. Amostra HG00262.....57

Figura 40 – Esquema sobre entropia de metilação. Figura extraída de https://nanoporetech.github.io/modkit/intro_entropy.html.....58

Figura 41 - Heatmap dos valores de entropia de metilação (ME) por gene/amostra.....59

Lista de tabelas

Tabela 1 - Genomas sequenciados via Oxford Nanopore. O genome HG002 foi utilizado para esta etapa. 12

Tabela 2 - Descrição das amostras utilizadas no estudo. A tabela apresenta informações sobre cada amostra do projeto 1000 Genomas sequenciada com a tecnologia Oxford Nanopore (ONT) R10, incluindo nome da amostra, sexo, código e nome da população de origem, além do código e nome da superpopulação correspondente. 12

Tabela 3 - Coordenadas genômicas dos marcadores analisados. A tabela apresenta as informações de cromossomo, posição inicial e final, identificador do marcador (ID) e o conjunto ao qual pertence. Os conjuntos incluem marcadores associados à cor dos olhos, genes de pigmentação, características capilares, SNPs do painel 34-plex, microhaplótipos para inferência de ancestralidade e marcadores relacionados à tonalidade da pele. 14

Tabela 4 - Tabelas de quantificação das amostras utilizando Qubit BR assay e Nanodrop. As linhas cinzas representam amostras com problemas durante as coletas, linhas vermelhas representam amostras com valores de A260/280 e A260/230 indicando baixa pureza no DNA. Amostras que apresentaram quantificações acima ou abaixo dos limites de detecção do Qubit foram marcadas em negrito e representam como 1000 ng/uL (Para concentrações acima) e 0 ng/uL (Para concentrações abaixo). 25

RESUMO

As tecnologias baseadas em DNA têm sido usadas na prática forense desde meados da década de 1980. Embora a genotipagem de STR baseada em PCR usando eletroforese capilar continue sendo o padrão ouro para a geração de perfis de DNA em casos de rotina em todo o mundo, a comunidade de pesquisa está continuamente buscando métodos alternativos capazes de fornecer informações adicionais para aumentar o poder de discriminação ou contribuir com novas pistas de investigação. A Oxford Nanopore Technologies (ONT) revolucionou o campo, oferecendo recursos em tempo real, resolução de molécula única e sequenciamento de leitura longa (LRS). A genética forense tem desempenhado um papel crucial na resolução de casos criminais e na identificação de indivíduos desconhecidos através da análise de DNA. Fenotipagem forense de DNA emergiu como uma área promissora, permitindo a previsão de características físicas de um indivíduo com base em seu perfil genético. A epigenética também ganhou destaque, explorando modificações químicas no DNA que podem influenciar a expressão gênica e fornecer informações adicionais sobre a identidade e características dos indivíduos. O objetivo geral deste estudo é implementar a metodologia de sequenciamento de genomas completos Oxford Nanopore e as estratégias bioinformáticas para análise destes dados. Os objetivos específicos incluem coordenar a coleta de amostras e extração de DNA a partir de tecido epitelial, realizar o processo de caracterização fenotípica das amostras coletadas, implementar a metodologia de sequenciamento Oxford Nanopore, conduzir análises computacionais para detecção de variantes e haplótipos, determinar as proporções de ancestralidade genômica e correlacioná-las com características de pigmentação humana. A coleta de tecido será realizada conforme protocolos estabelecidos para preservação da integridade do DNA. O isolamento de DNA, preparação de biblioteca, sequenciamento e identificação de bases serão conduzidos seguindo as diretrizes da Oxford Nanopore Technologies. O pipeline de alinhamento e identificação de variantes será implementado utilizando ferramentas bioinformáticas adequadas. A análise de metilação, predição de idade e pigmentação nas amostras serão realizadas com base em modelos estatísticos previamente validados.

Palavras-chaves: Genética forense; Sequenciamento de leitura longa; Fenotipagem forense de DNA; Tecnologias Oxford Nanopore; Epigenética.

1. INTRODUÇÃO

Desde meados da década de 1980, várias tecnologias baseadas em DNA têm sido aplicadas na prática forense para solucionar crimes. A progressão do sequenciamento do tipo Sanger para o sequenciamento paralelo maciço abriu novos caminhos para uma análise de DNA rápida, eficiente e econômica, trazendo novas possibilidades para o campo forense. Embora o método padrão ouro para genotipagem forense continue sendo a análise de polimorfismos STR por meio de eletroforese capilar, os laboratórios forenses estão introduzindo gradualmente ensaios complementares baseados em plataformas tecnológicas mais avançadas. O esforço original de sequenciamento do genoma humano exigiu pelo menos 13 anos para ser realizado e um investimento de US$ 3 bilhões. Em contraste, a tecnologia atual permite o sequenciamento do genoma humano por menos de US$ 1.000 em poucas horas. No início dos anos 2000, as plataformas de sequenciamento ocupavam um armazém inteiro, enquanto hoje as menores plataformas são do tamanho de um smartphone. Esta análise explora a evolução histórica dos métodos de sequenciamento de ácidos nucleicos e suas aplicações na ciência forense, com atenção especial ao sequenciamento de leitura longa (LRS). de alto rendimento.

1.1 Genética forense

A genética forense progrediu junto com os avanços impressionantes nas tecnologias de análise de DNA. A implementação de tecnologias baseadas em sequenciamento de leituras curtas (SRS) trouxe avanços consideráveis para o campo, permitindo uma melhor exploração da variabilidade do genoma devido à sua capacidade de sequenciar várias amostras e centenas de milhares de marcadores genéticos diferentes, como *Short Tandem Repeat* (STR) e Polimorfismo de Nucleotídeo Único (SNP[1]), simultaneamente, a partir de pequenas quantidades de DNA (Carratto et al., 2022). Ainda assim, o alto custo de implementação do SRS e suas leituras de tamanho curto podem resultar em limitações para a análise de DNA em alguns cenários, especialmente para fins forenses. As metodologias de LRS superam essas restrições, com suas leituras longas e implementação de custo relativamente baixo, pelo menos para Oxford Nanopore Tecnologies (ONT), além de seu sequenciamento em tempo real, de molécula única e de alta velocidade. Embora o Illumina SRS tenha sido usado em ensaios sem reação em cadeia de polimerase (PCR) (Byrska-Bishop et al., 2022), isso não é comum, e a ONT também permite a possibilidade de preparar bibliotecas a partir de amostras de DNA nativo, eliminando a etapa de amplificação e os artefatos que podem ser gerados na etapa de PCR. Os benefícios não acabam aqui: O LRS pode ser usado para analisar o status de metilação do DNA enquanto se analisa a sequência de DNA sem a necessidade de conversão prévia em bissulfito. Além disso, a Oxford

[1]Do inglês *Single Nucleotide Polymorphism*.

Nanopore também analisa fitas de RNA nativas sem a necessidade de conversão prévia em DNA complementar (cDNA). No entanto, a LRS ainda não foi amplamente explorada na área forense, provavelmente devido às suas taxas de erro relativamente altas e à suposição generalizada de que ela sempre exige uma quantidade maior de DNA de entrada em comparação com os kits forenses já validados para uso de SRS. Mesmo assim, a LRS tem se aprimorado constantemente, e o número de artigos publicados empregando essa metodologia para fins forenses tem crescido anualmente. Até o momento, a ONT tem sido mais usada do que a Pacific Biosciences (PacBio) na área forense: uma rápida pesquisa no Pubmed com os termos *Oxford Nanopore + forensics* e *PacBio + forensics* retornou 31 correspondências para a primeira e apenas 12 para a segunda. A PacBio é mais cara que a ONT e geralmente requer instalações dedicadas, o que dificulta sua implementação em ambientes forenses. No entanto, a PacBio é uma tecnologia altamente precisa com suas leituras de sequenciamento de consenso circular (CCS) e permite uma análise poderosa do DNA mitocondrial, transcriptômica e metagenômica, que podem ser úteis para fins forenses (**Figura 1**).

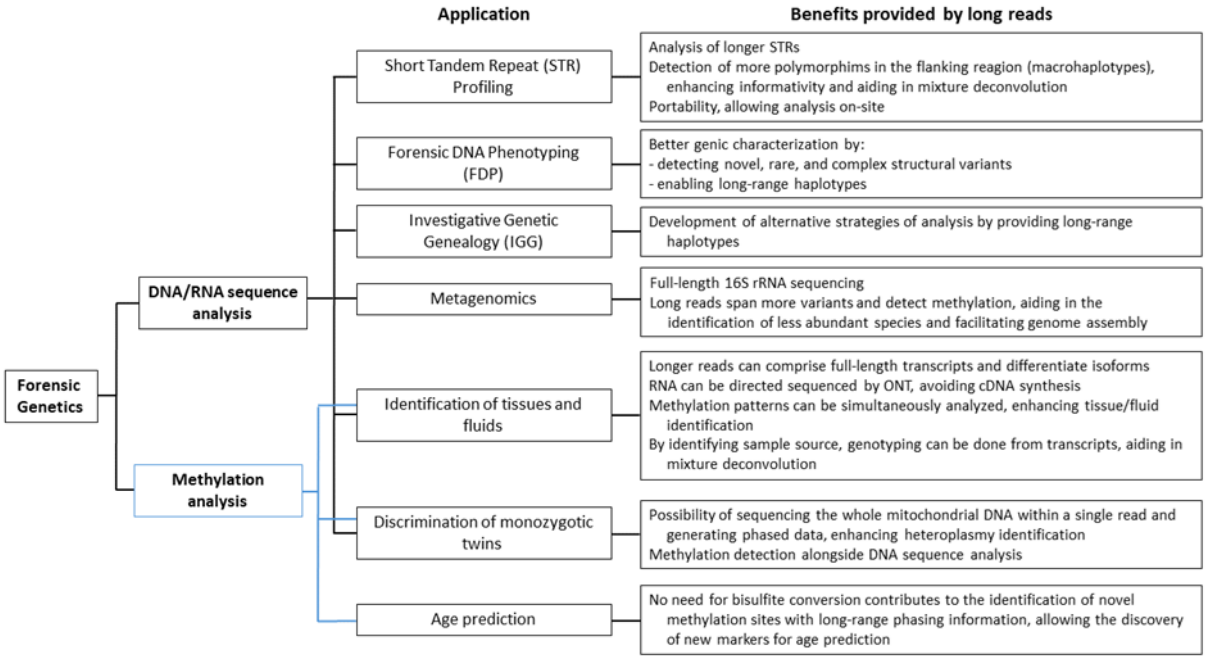


Figura 1 - Principais benefícios proporcionados pelo sequenciamento de leitura longa para a genética forense. Figura republicada de (Ferreira et al., 2025).

1.1 Fenotipagem forense de DNA

Embora o perfil de STR seja a análise padrão ouro em casos forenses, ele pode não ser imediatamente útil em situações em que o perfil gerado não corresponda a nenhum suspeito ou a nenhum outro perfil em bancos de dados. Além disso, os STRs têm sequências mais longas e, em casos de amostras degradadas, podem ser difíceis de amplificar, gerando perfis parciais. Assim, tornou-se de suma importância a avaliação de outros tipos de polimorfismos que pudessem fornecer contribuições adicionais para as investigações forenses. Nesse contexto, os SNPs

aparecem como uma boa opção, pois são mutações de base única no DNA, exigindo amplicons mais curtos e sendo mais fáceis de analisar do que os STRs. Embora vários SNPs precisem ser analisados para se obter um poder de discriminação comparável, isso não é mais uma limitação com o advento do sequenciamento massivo paralelo (MPS).

No final dos anos 2000, surgiu um novo campo de estudo chamado Fenotipagem Forense de DNA (FDP). Seu objetivo é prever características externamente visíveis (EVCs) de um indivíduo analisando apenas SNPs-alvo (e, às vezes, InDels) em seu DNA. Assim, sempre que o perfil tradicional de STR falhar, a FDP poderá gerar novas pistas sobre a aparência do indivíduo, ajudando nas investigações policiais.

Entre os EVCs, a pigmentação foi a primeira a desenvolver e validar ferramentas de FDP. O sistema IrisPlex (Walsh et al., 2011) foi publicado pela primeira vez em 2011, compreendendo seis SNPs para a previsão da cor dos olhos. Modelos para cabelo e pele foram desenvolvidos posteriormente, integrando o sistema HIrisPlex-S (Breslin et al., 2019; Chaitanya et al., 2018), capaz de prever simultaneamente a cor dos olhos, do cabelo e da pele a partir de 41 marcadores (40 SNPs e 1 InDel). Inicialmente, dois ensaios de mini-sequenciamento (SNaPshot) foram projetados para análise de marcadores HIrisPlex-S em eletroforese capilar (CE).

No entanto, como a técnica FDP ainda está em desenvolvimento, o uso da tecnologia SRS traz vantagens quando comparada à CE. Os ensaios SNaPshot têm baixa capacidade multiplex, portanto, são necessários cada vez mais ensaios para acomodar novos marcadores de FDP, o que significa que também é necessário mais DNA de entrada. Por outro lado, o SRS pode analisar milhares de marcadores de uma só vez, e a incorporação de novos polimorfismos em seus ensaios é muito mais fácil. Isso pode permitir o aumento da previsão de precisão, já que centenas ou milhares de marcadores podem ser incluídos em um único ensaio, e pode permitir a avaliação simultânea de dezenas de EVCs. Portanto, os cientistas da FDP decidiram direcionar seus esforços para a implementação de ferramentas da FDP em plataformas SRS. Não demorou muito para que os desenvolvedores do HIrisPlex-S propusessem um protocolo para gerar e analisar os 41 marcadores a partir de dados brutos de SRS (Breslin et al., 2019). Os kits de preparação de amostras para análise de SRS visando os marcadores 24 (para olhos e cabelos) e 41 (para olhos, cabelos e pele) já estão disponíveis comercialmente.

Um evento notável no campo de FDP foi a consolidação do Consórcio *Visible Attributes Through Genomics* (VISAGE), uma parceria entre oito países europeus que visa aprimorar a tecnologia de FDP por meio do desenvolvimento de novas ferramentas. Eles desenvolveram uma ferramenta aprimorada com mais de 500 marcadores para inferir a ancestralidade e prever a cor dos olhos, do cabelo, das sobrancelhas e da pele, o formato do cabelo, a calvície masculina e a presença de sardas. Os marcadores para cor dos olhos, do cabelo e da pele são os mesmos do

HIrisPlex-S. Devido ao grande número de marcadores, o projeto foi totalmente desenvolvido e validado em plataformas SRS. No entanto, faltam estudos adicionais que explorem o poder de previsão dos marcadores VISAGE em diferentes populações.

Embora muito tenha sido feito desde a primeira concepção do FDP, ainda há espaço para melhorias. O HIrisPlex-S é uma ferramenta bem estabelecida, mas ainda apresenta dificuldades na previsão de alguns fenótipos, como cores intermediárias de olhos e cabelos castanhos e tons de pele em geral (Carratto et al., 2019, 2021). A previsão de outras características nos modelos VISAGE também pode ser aprimorada. Além disso, outras características, como a altura do corpo e o formato do rosto, são altamente poligênicas e continuam sendo um desafio para os cientistas do FDP.

Nesse sentido, o LRS pode ser benéfico para a fenotipagem de DNA. Foi comprovado que a análise de SNP é mais fácil usando essas tecnologias em comparação com STRs (Cornelis et al., 2017; Ren et al., 2021; Tytgat et al., 2022), mesmo para SNPs tri-alélicos (Cornelis et al., 2019). As leituras longas podem fornecer uma melhor caracterização dos genes envolvidos no surgimento de fenótipos, detectando variantes de estrutura novas, raras e complexas (Ling et al., 2023). Como resultado, haplótipos completos de longo alcance (incluindo marcadores epigenéticos, conforme discutido abaixo) de genes inteiros que influenciam características físicas podem ser identificados e empregados para previsão forense. Isso também permite uma exploração mais profunda do papel dos haplótipos e da epistasia no surgimento de fenótipos, conforme sugerido por estudos anteriores de sua influência nas EVCs (Pośpiech et al., 2014; Ruiz et al., 2013).

1.2 Epigenética

A epigenética desempenha um papel fundamental na regulação da expressão gênica. Entre os principais mecanismos epigenéticos está a metilação do DNA (5mC), que é um dos quatro tipos existentes de modificação da citosina (**Figura 2**). A metilação do DNA é geralmente analisada por duas abordagens diferentes: matriz e sequenciamento. O *HumanMethylation450 BeadChip* da Illumina, composto por mais de 450 mil marcadores, costumava ser uma escolha popular, mas foi substituído pelo *MethylationEPIC BeadChip*, que tem como alvo mais de 935 mil sítios CpG em sua nova versão. As tecnologias de sequenciamento comumente usadas incluem pirosequenciamento, SNaPshot ou SRS. No entanto, a metilação não pode ser detectada a partir da análise padrão de DNA, portanto, independentemente da metodologia escolhida, é necessária uma etapa prévia de conversão de bissulfito. Essa técnica requer uma grande quantidade de DNA de entrada e causa danos ao DNA, o que não é ideal para a ciência forense, pois amostras de baixa quantidade/qualidade são comuns nesse campo. Nesse sentido, a análise de metilação pode se beneficiar do LRS, pois essa abordagem pode detectar a metilação a partir

de protocolos de sequenciamento de DNA padrão, evitando danos adicionais à amostra e a possível tendência que pode ser introduzida durante a etapa de conversão de bissulfito.

A conversão de bissulfito também resulta em perda de informações, pois duas modificações de DNA (5fC e 5caC) são convertidas em timina juntamente com citosina, enquanto as outras duas (5mC e 5hmC) permanecem como citosinas (**Figura 2**) (Searle et al., 2023). A ONT demonstrou ser capaz de diferenciar todas as quatro modificações em cadeias sintetizadas (Wescoe et al., 2014). O PacBio pode detectar melhor as modificações com grupos -OH, pois elas causam uma maior interferência de sinal. Portanto, as modificações que não apresentam um grupo hidroxila, como a metilação, podem não ser detectadas pela PacBio. Uma possível solução para melhorar isso seria o uso de técnicas de conversão alternativas que não danificam o DNA e aumentam a perturbação do sinal (por exemplo, adicionando grupos hidroxila a essas modificações) (Searle et al., 2023). Além disso, a melhoria contínua da precisão da tecnologia e o desenvolvimento de ferramentas bioinformáticas sofisticadas capazes de detectar modificações de sinal e traduzi-las em modificações específicas do DNA são fundamentais para as tecnologias PacBio e ONT. Nesse sentido, foram propostos chamadores de metilação de haplótipos (Kolmogorov et al., 2023; Ni et al., 2023).

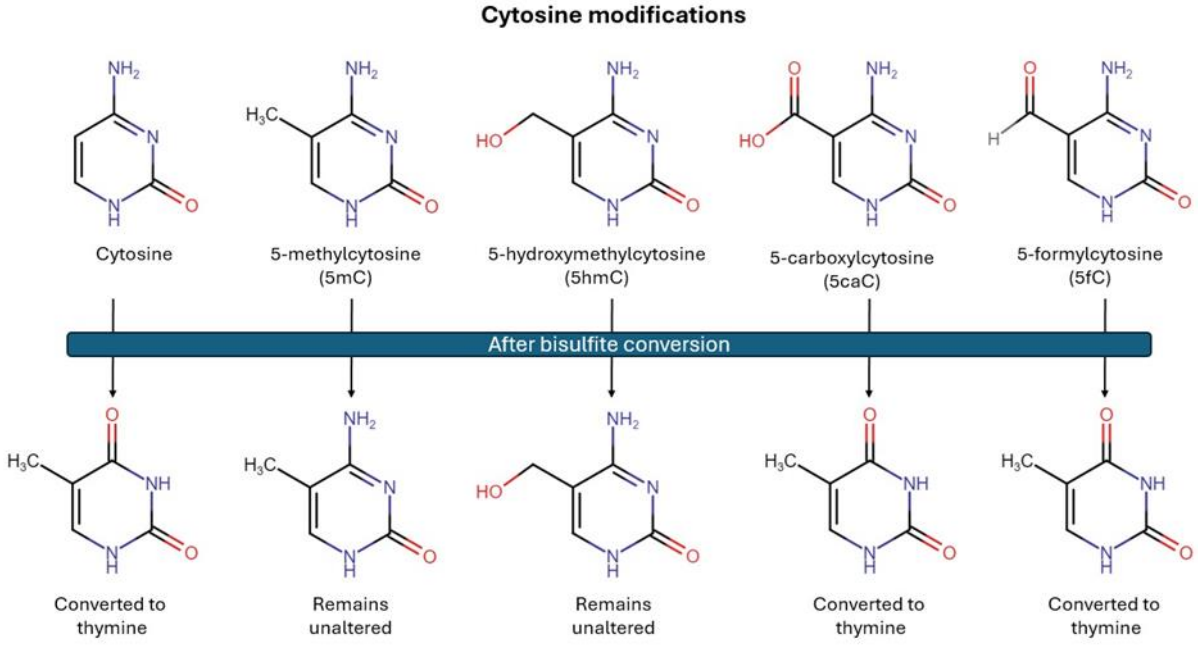


Figura 2 - Modificações de citosina antes e depois da conversão de bissulfito. Figura retirada de (Ferreira et al., 2025).

Na genética forense, a análise de metilação é aplicada principalmente em três subáreas: identificação de fluidos e tecidos, diferenciação de gêmeos monozigótico (MZ) e previsão de idade (Vidaki & Kayser, 2018). Como cada fluido e tecido apresenta seu próprio padrão de expressão gênica e a metilação desempenha um papel regulador fundamental nesse processo, os padrões de metilação variam entre os tecidos/fluidos. Esses padrões também são influenciados

por fatores de estilo de vida, como exposição ao sol, dieta, consumo de álcool e tabagismo, e, portanto, gêmeos MZ, que basicamente têm o mesmo DNA nuclear, podem apresentar perfis de metilação diferentes. Outro fator que influencia a metilação é a idade. Além do estilo de vida, algumas regiões do DNA são especificamente metiladas ou não metiladas ao longo do tempo, e muitos painéis foram propostos para a previsão da idade cronológica (Aliferi et al., 2022; Freire-Aradas et al., 2022; Woźniak et al., 2021).

A metilação tem sido explorada para enfrentar os desafios nessas três áreas principais. No entanto, ainda há espaço para melhorias. As leituras longas têm o potencial de aprimorar essas análises, detectando a metilação em regiões que podem ser danificadas durante a conversão de bissulfito, sondando regiões não visadas nas matrizes e fornecendo sequências longas em fases que incluem locais de variação e modificações epigenéticas. Esse aprimoramento poderia contribuir para a identificação de novos marcadores preditivos para o desenvolvimento de novos painéis para identificação de fluidos/tecidos e previsão de idade. Além disso, ele facilita uma exploração mais profunda dos locais de metilação, auxiliando na identificação de diferenças entre os epigenomas de dois gêmeos monozigóticos.

1.3 Sequenciamento do genoma completo via Oxford Nanopore Technologies

No panorama do sequenciamento de terceira geração, a Oxford Nanopore Technologies (ONT) se destaca como uma vanguarda, contribuindo com inovações distintas para o cenário do sequenciamento de DNA/RNA. Pioneira no conceito de sequenciamento de nanoporos, a ONT emprega nanoporos biológicos incorporados em membranas sintéticas para analisar ácidos nucleicos em tempo real (Jain et al., 2018; Y. Wang et al., 2021). O mecanismo de sequenciamento baseado em nanoporos permite a detecção eletrônica direta de sequências de nucleotídeos à medida que elas passam pelo nanoporo, oferecendo vantagens em relação ao comprimento de leitura e à adaptabilidade a vários tipos de amostras (**Figura 3**). Por exemplo, isso permite que a fita de RNA seja sequenciada diretamente, ou seja, sem sintetizar DNA complementar, facilitando a identificação de isoformas e a determinação do comprimento da cauda poli-A (Wan et al., 2022; J. Wang et al., 2024; Y. Wang et al., 2021; Workman et al., 2019). Ignorando etapas indutoras de viés, como a PCR, o sequenciamento direto fornece dados precisos e nativos sobre a molécula alvo (Bowden et al., 2019; Delahaye & Nicolas, 2021; Kono & Arakawa, 2019).

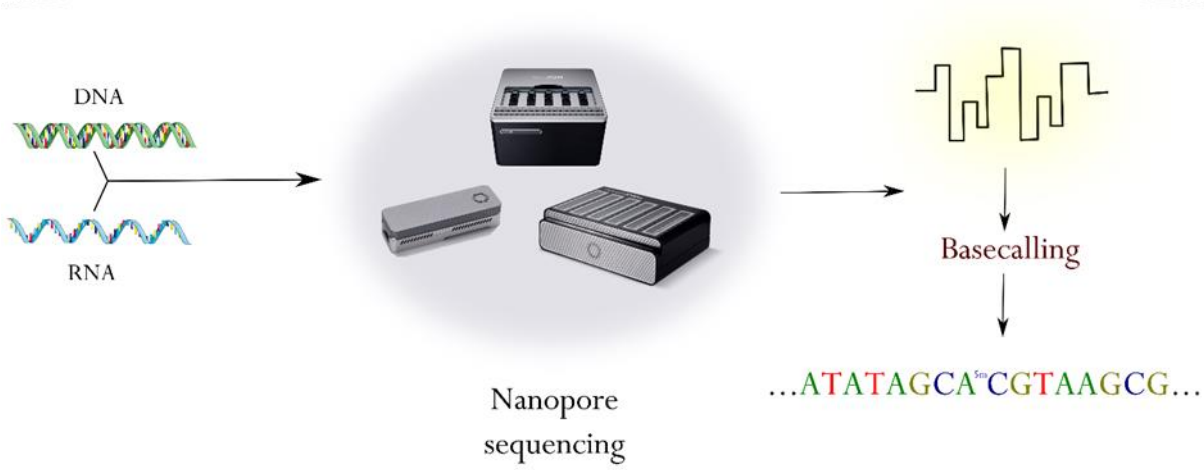


Figura 3 - Visão geral do sequenciamento Oxford Nanopore. Os elementos gráficos foram obtidos em https://smart.servier.com/, enquanto as imagens do equipamento Oxford Nanopore foram baixadas de https://nanoporetech.com/.

O sequenciamento por nanoporos tem dois modos: simplex e duplex. No modo simplex (também conhecido como 1D), um nanoporo sequência apenas uma fita de DNA. No modo duplex, ambas as fitas de DNA são sequenciadas (uma imediatamente após a outra), resultando em um procedimento de sequenciamento duplo que facilita a correção de bases (B. Lin et al., 2021; Nanoporetech, 2017; Y. Wang et al., 2021). ONTs geralmente produzem leituras na faixa de 10 a 100 kb, uma diferença substancial em relação às leituras dos sequenciadores Illumina, com leituras em torno de 150 bases (Logsdon et al., 2020). Fragmentos mais longos do que esses são frequentemente obtidos e classificados como leituras ultralongas, com a leitura mais longa registrada em um tamanho de 4,2 Mb usando um kit (SQK-ULK114) desenvolvido para otimizar a produção de sequências muito longas (Jain, 2021).

Como mencionado anteriormente, a versatilidade da tecnologia de sequenciamento de nanoporos se estende notavelmente ao campo da epigenética, apresentando uma abordagem pioneira para a análise de modificações químicas de DNA e RNA em escala genômica (Wan et al., 2022; Y. Wang et al., 2021). Quando a sequência de bases contém modificações químicas, como metilação ou hidroximetilação, por exemplo, a ONT pode detectar um padrão de sinal elétrico diferente dos não modificados. Dessa forma, a detecção de metilação, por exemplo, é obtida evitando-se o uso de tratamentos químicos/enzimáticos e a necessidade de PCR, que pode incluir vieses na análise (Y. Wang et al., 2021). O método de sequenciamento por nanoporos tem o potencial de detectar qualquer modificação de nucleotídeo na sequência. Ainda assim, há modelos pré-treinados para as modificações mais comuns, como 5mC, 5hmC e 6mA (Gershman et al., 2022; Liu et al., 2021; Nanoporetech, 2023d; Y. Wang et al., 2021). Outras modificações exigirão que o usuário treine seu modelo. A ONT recomenda uma profundidade de cobertura de 20x para a análise de 5mC e 5hmC e uma cobertura de cerca de 30x para a faseamento de dados.

1.4 Justificativa para este projeto

Este trabalho explorou o potencial da epigenética e do sequenciamento de leitura longa (Long-Read Sequencing, LRS) na genética forense, com foco na fenotipagem e na predição de características individuais a partir de amostras de tecido epitelial. A introdução de tecnologias avançadas, como o sequenciamento por nanoporos da Oxford Nanopore Technologies (ONT), proporcionou uma oportunidade única para aprimorar a análise forense, permitindo uma visão mais abrangente e precisa do genoma humano.

O avanço das técnicas de genotipagem forense baseadas em DNA, desde o sequenciamento de Sanger até o Massively Parallel Sequencing (MPS), revolucionou a investigação criminal. No entanto, evidenciou-se a necessidade de métodos complementares capazes de superar as limitações das abordagens convencionais. O LRS apresentou vantagens significativas, como a capacidade de analisar não apenas a sequência de DNA, mas também os padrões de metilação epigenética, sem a necessidade de conversão prévia em bissulfito. Isso mostrou-se particularmente relevante para a genética forense, onde amostras de baixa quantidade e qualidade são comuns, tornando os métodos tradicionais de análise de metilação menos viáveis.

FDP é uma área emergente que busca prever características físicas individuais a partir do perfil genético. A análise de SNPs tem sido fundamental nesse contexto, e o desenvolvimento de novas ferramentas, como o sistema HIrisPlex-S, demonstrou a viabilidade dessa abordagem. No entanto, ainda existem desafios, como a precisão na previsão de fenótipos intermediários e a incorporação de informações epigenéticas para uma análise mais completa.

A epigenética, em particular a metilação do DNA, desempenha um papel crucial na regulaçãoda expressão gênica e na determinação de características fenotípicas. A análise de padrões de metilação pode fornecer informações valiosas sobre a idade e a pigmentação da pele de um indivíduo, complementando os dados genéticos tradicionais. O uso do LRS permite uma análise mais abrangente dos padrões de metilação, evitando os artefatos introduzidos pela conversão de bissulfito e fornecendo uma visão mais precisa dos locais de metilação em todo o genoma. Além disso, o sequenciamento de leitura longa tem o potencial de facilitar a diferenciação entre diferentes fluidos e tecidos, bem como a previsão de características individuais com base em modificações epigenéticas específicas.

Dessa forma, este projeto explorou as aplicações do LRS e da análise de metilação epigenética na genética forense, com ênfase na fenotipagem e na predição de características individuais a partir de tecido epitelial. A integração dessas tecnologias avançadas demonstrou potencial para fornecer uma compreensão mais completa e precisa do perfil genético de um indivíduo, contribuindo para investigações criminais mais eficazes e justas.

2. OBJETIVOS

O objetivo geral consistiu na implementação da metodologia de sequenciamento de genomas completos Oxford Nanopore e das estratégias bioinformáticas para análise destes dados para predição de fenótipos.

Os objetivos específicos desta projeto foram:

- Coordenar a coleta de amostras e extração de DNA a partir de tecido epitelial;
- Realizar o processo de caracterização fenotípica das amostras coletadas;
- Implementar a metodologia de sequenciamento Oxford Nanopore;
- Conduzir análise computacional para detecção das variantes e haplótipos dos genomas sequenciados;

3. MATERIAIS E MÉTODOS

A seção de materiais e métodos será dividida em duas partes. A primeira sobre os procedimentos para coleta de amostras e extração de DNA dos diferentes tecidos. Já a segunda sobre os procedimentos computacionais para análises.

3.1 Parte 1 – Coleta de amostras e determinação de métodos de extração de DNA

3.1.1 Coleta de amostras e caracterização fenotípica da amostra populacional

A coleta foi coordenada junto aos professores Dr. Hélio Amante Miott e Dra. Anna Carolina Miola (FMB-UNESP), no Ambulatório de Dermatologia do Hospital das Clínicas da Faculdade de Medicina de Botucatu (HCFMB), da Unesp. Ao todo, foram coletadas amostras de 32 indivíduos. O projeto encontra-se aprovado pelo comitê de ética com o seguinte CAAE: 71548223.8.0000.5411. Foram coletadas amostras de pele, sangue e saliva de cada indivíduo. Aproximadamente 5 mL de sangue periférico foram coletados por punção venosa, utilizando tubos com anticoagulante EDTA. O sangue foi imediatamente armazenado em temperatura controlada (4 °C) até o processamento. Foram coletados aproximadamente 2 mL de saliva utilizando tubos de coleta GeneFiX Saliva DNA/RNA Collection & Stabilization (Isohelix, Cell Projects, Kent, UK), e as amostras foram armazenadas imediatamente a 4 °C até a etapa de extração. Uma biópsia de pele de aproximadamente 500 mg foi obtida utilizando *punch* cirúrgico de 4 mm. O tecido foi imediatamente armazenado em solução de RNA Later (Ambion), para preservação de ácidos nucleicos, e mantido a -80 °C até o processamento. O tecido obtido foi dividido em duas partes iguais, reservadas para extração de DNA e RNA.

Durante a coleta das amostras, os indivíduos foram submetidos a um questionário sobre características étnicas e morfológicas, além da coleta de fotos da pele nas regiões da testa e braço, olhos, raízes e mechas do cabelo. Por fim, um colorímetro Colaris Color Muse (Colaris) foi utilizado para realizar medidas fotoespectrofotométricas da pele nas regiões da testa/nariz, braço interno e dorso da mão esquerda, para obtenção de valores L*, a* e b*. As leituras foram realizadas em triplicatas.

3.1.2 Extração de DNA

O DNA de alto peso molecular foi extraído utilizando protocolos específicos para cada tipo de tecido. A extração de DNA para pele foi realizada utilizando um protocolo adaptado com o kit Blood & Cell Culture DNA Midi Kit (Qiagen). Inicialmente, 19 µL de RNase A, 100 mg/mL, (Qiagen) foram adicionados a um tubo Falcon de 15 mL contendo 9,5 mL de tampão G2. Em seguida, até 100 mg de tecido de pele foram triturados utilizando pinças e bisturi, e o material resultante foi transferido para o tubo contendo o tampão G2. Posteriormente, 250 µL de proteinase K (Qiagen) foram adicionados, e o tubo foi misturado vigorosamente em vórtex para garantir a homogeneidade da suspensão, essencial para uma lise eficiente. A suspensão foi incubada a 56 °C durante a noite. Após a incubação, caso o lisado apresentasse heterogeneidade ou pedaços visíveis de tecido, o tubo era centrifugado a 2000×g por 10 minutos a 4°C. O pellet foi descartado, e o sobrenadante homogêneo foi cuidadosamente transferido para um novo recipiente. Esta etapa foi fundamental, pois lisados não homogêneos poderiam obstruir a coluna Genomic-tip. O lisado foi então purificado utilizando o protocolo Genomic-tip da QIAGEN (passos 1–6, conforme descrito nas páginas 49–52 do manual do fabricante). Para otimizar o rendimento de DNA, a eluição foi realizada por 2 horas a 50°C, utilizando 150 µL de TE (1 mM EDTA, pH 8,0), com mistura ocasional do conteúdo do tubo por inversão suave.

A extração de DNA a partir de amostras de sangue e saliva foi realizada utilizando o kit Puregene Blood Kit (Qiagen). Para as amostras de sangue, 3 mL foram transferidos para um tubo Falcon de 15 mL contendo 9 mL de solução de lise RBC (RBC Lysis Solution). O conteúdo foi homogeneizado por inversão por 10 vezes, seguido de incubação à temperatura ambiente por 5 minutos. Após esse período, as amostras foram centrifugadas a 2.000×g por 2 minutos. O sobrenadante foi cuidadosamente descartado, deixando aproximadamente 200 µL de líquido residual para preservar o pellet de células brancas. Em seguida, o tubo contendo o pellet foi submetido a vórtex para ressuspensão completa. Foram adicionados 3 mL de solução de lise celular (Cell Lysis Solution), e a amostra foi novamente submetida a vórtex por 10 segundos para garantir a lise total das células. Para digestão de RNA, foram adicionados 15 µL de RNase A (100 mg/mL), e o tubo foi incubado a 37 °C por 15 minutos. Após a digestão, as amostras foram refrigeradas em gelo por 3 minutos, estabilizando o DNA. Posteriormente, foi adicionado 1 mL de solução de precipitação de proteína (Protein Precipitation Solution), e o conteúdo foi

homogeneizado vigorosamente por vórtex. As amostras foram centrifugadas a 2.000×g por 5 minutos, e o sobrenadante resultante foi cuidadosamente transferido para um novo tubo. A precipitação do DNA foi realizada adicionando-se 3 mL de isopropanol gelado ao sobrenadante. O tubo foi invertido 10 vezes para promover a formação do precipitado de DNA e então centrifugado a 2.000×g por 3 minutos. O pellet de DNA obtido foi lavado com 3 mL de etanol 70% gelado e centrifugado novamente a 2.000×g por 1 minuto. Após a lavagem, o pellet foi seco em temperatura ambiente por 10 minutos e dissolvido em 300 µL de solução de hidratação de DNA (DNA Hydration Solution). Finalmente, as amostras foram incubadas a 65 °C por 1 hora para garantir a dissolução completa do DNA.

Para a extração de DNA de amostras de saliva, 2 mL de saliva foram misturados a 5,5 mL de solução de lise celular (Cell Lysis Solution), 30 µL de Proteinase K (Qiagen) e 30 µL de RNase A (100 mg/mL, Qiagen). A mistura foi incubada a 56 °C por 1 hora para garantir a lise eficiente das células e a digestão de proteínas e RNA. Após a incubação, foram adicionados 2 mL de solução de precipitação de proteínas (Protein Precipitation Solution), seguidos de agitação vigorosa em vórtex por 20 segundos. A amostra foi então incubada em gelo por 5 minutos para favorecer a precipitação das proteínas. Posteriormente, o tubo foi centrifugado a 2.000×g por 10 minutos. O sobrenadante resultante foi cuidadosamente transferido para um novo tubo, evitando qualquer resíduo do pellet proteico. Para precipitar o DNA, foram adicionados 6 mL de isopropanol gelado ao sobrenadante, e o tubo foi invertido 10 vezes para promover a formação do precipitado de DNA. Em seguida, as amostras foram centrifugadas a 2.000×g por 10 minutos. O pellet de DNA formado foi lavado com 6 mL de etanol 70% gelado e novamente centrifugado a 2.000×g por 1 minuto. Após a lavagem, o pellet foi seco em temperatura ambiente por 10 minutos e dissolvido em 200 µL de solução de hidratação de DNA (DNA Hydration Solution). Para garantir a completa dissolução do DNA, as amostras foram incubadas a 65 °C por 1 hora.

O DNA extraído foi quantificado em um Fluorômetro Qubit (Invitrogen) usando o ensaio de espectro amplo (BR assay) para dsDNA. A qualidade do DNA para sequenciamento foi avaliada usando um Espectrofotômetro NanoDrop (ThermoFisher) após a quantificação.

3.2 Análises computacionais

3.2.1 Dados públicos

Foram utilizadas duas fontes de amostras no presente estudo. realizamos a análise de dados públicos provenientes do projeto 1000 Genomas, especificamente de amostras sequenciadas utilizando a tecnologia Oxford Nanopore (ONT) com a química R10. Ao todo, foram incluídas 57 amostras, das quais 52 foram geradas por meio de protocolos de *long reads* e 5 por *ultra long reads*. Esses dados foram obtidos a partir do repositório público disponível em

1000G ONT (https://s3.amazonaws.com/1000g-ont/index.html?prefix=), garantindo acesso a sequências de alta qualidade para nossas análises subsequentes.

A outra fonte foi uma amostra disponibilizada pela Oxford Nanopore através do sistema Amazon Web Services (AWS) de sequenciamentos de amostras do genome in a bottle (GIAB) (https://www.nist.gov/programs-projects/genome-bottle). São disponibilizadas três amostras em trio, pais e filho, das linhagens (Genome in a Bottle Ashkenazi Trio) totalizando quase 8 Tb de dados (arquivos de sequenciamento e outras análises) (**Tabela 1**). Os códigos de acesso são descritos como:

Tabela 1 - Genomas sequenciados via Oxford Nanopore. O genome HG002 foi utilizado para esta etapa.

Código Genoma	Relação	Linhagem celular (Coriell)
HG002	Filho	GM24385
HG003	Pai	GM24149
HG004	Mãe	GM24143

Estas amostras foram sequenciadas em duas *flowcells* R10.4.1 do PromethION gerando aproximadamente 200 Gbases (Refletindo em uma cobertura de cerca de 60x). O kit de ligação versão 14 (https://store.nanoporetech.com/ligation-sequencing-kit-v14.htm) foi utilizado para o preparo destas amostras. Devido a limitação momentânea de espaço realizamos o download apenas da amostra HG002 através do *bucket* S3 (s3://ont-open-data/giab_2023.05/flowcells/hg002/20230424_1302_3H_PAO89685_2264ba8c/pod5_pass/) utilizando aws-cli. Foram baixados todos os arquivos POD5 (formato atual de saída de sequenciadores ONT) totalizando 1.3 Tb de dados. Utilizamos esta amostra para verificar os procedimentos de análise de metilação e de identificação de variantes.

3.2.2 Amostras utilizadas

A **Tabela 2** apresenta informações detalhadas sobre cada amostra, incluindo nome, sexo, código e nome da população de origem, bem como a superpopulação à qual pertencem. Esses metadados são essenciais para contextualizar as análises realizadas e interpretar variações genéticas em diferentes grupos populacionais.

Tabela 2 - Descrição das amostras utilizadas no estudo. A tabela apresenta informações sobre cada amostra do projeto 1000 Genomas sequenciada com a tecnologia Oxford Nanopore (ONT) R10, incluindo nome da amostra, sexo, código e nome da população de origem, além do código e nome da superpopulação correspondente.

Sample name	Sex	Population code	Population name	Superpopulation code	Superpopulation name
GM19238	female	YRI	Yoruba in Ibadan, Nigeria	AFR	African Ancestry
GM19429	male	LWK	Luhya in Webuye, Kenya	AFR	African Ancestry

GM20847	female	GIH	Gujarati Indians in Houston, Texas, USA	AMR	American Ancestry
HG00097	female	GBR	British	EUR	European Ancestry
HG00102	female	GBR	British	EUR	European Ancestry
HG00109	male	GBR	British	EUR	European Ancestry
HG00142	male	GBR	British	EUR	European Ancestry
HG00143	male	GBR	British	EUR	European Ancestry
HG00145	male	GBR	British	EUR	European Ancestry
HG00171	female	FIN	Finnish	EUR	European Ancestry
HG00233	female	GBR	British	EUR	European Ancestry
HG00237	female	GBR	British	EUR	European Ancestry
HG00262	female	GBR	British	EUR	European Ancestry
HG00263	female	GBR	British	EUR	European Ancestry
HG00277	male	FIN	Finnish	EUR	European Ancestry
HG00326	female	FIN	Finnish	EUR	European Ancestry
HG00372	male	FIN	Finnish	EUR	European Ancestry
HG00463	male	CHS	Southern Han Chinese	EAS	East Asian Ancestry
HG00729	female	CHS	Southern Han Chinese	EAS	East Asian Ancestry
HG01063	male	PUR	Puerto Rican	AMR	American Ancestry
HG01258	male	CLM	Colombian	AMR	American Ancestry
HG02312	female	PEL	Peruvian	AMR	American Ancestry
HG02615	female	GWD	Gambian Mandinka	AFR	African Ancestry
HG02763	female	GWD	Gambian Mandinka	AFR	African Ancestry
HG02772	female	GWD	Gambian Mandinka	AFR	African Ancestry
HG02861	female	GWD	Gambian Mandinka	AFR	African Ancestry
HG02884	male	GWD	Gambian Mandinka	AFR	African Ancestry
HG02885	female	GWD	Gambian Mandinka	AFR	African Ancestry
HG02887	male	GWD	Gambian Mandinka	AFR	African Ancestry
HG02891	female	GWD	Gambian Mandinka	AFR	African Ancestry
HG02952	female	ESN	Esan	AFR	African Ancestry
HG02968	male	ESN	Esan	AFR	African Ancestry
HG02977	male	ESN	Esan	AFR	African Ancestry
HG03009	male	BEB	Bengali	SAS	South Asian Ancestry
HG03160	male	ESN	Esan	AFR	African Ancestry
HG03166	male	ESN	Esan	AFR	African Ancestry
HG03352	male	ESN	Esan	AFR	African Ancestry
HG03367	male	ESN	Esan	AFR	African Ancestry
HG03436	male	MSL	Mende	AFR	African Ancestry
HG03575	female	MSL	Mende	AFR	African Ancestry
HG03583	female	MSL	Mende	AFR	African Ancestry
HG03629	male	PJL	Punjabi	SAS	South Asian Ancestry
HG03694	male	STU	Tamil	SAS	South Asian Ancestry
HG03745	male	STU	Tamil	SAS	South Asian Ancestry
HG03752	female	STU	Tamil	SAS	South Asian Ancestry
HG03755	male	STU	Tamil	SAS	South Asian Ancestry
HG03772	female	ITU	Telugu	SAS	South Asian Ancestry
HG04054	female	ITU	Telugu	SAS	South Asian Ancestry

3.2.3 *Coordendas genômicas*

As coordenadas genômicas utilizadas neste estudo correspondem a variantes e regiões associadas a características fenotípicas e ancestrais. A **Tabela 3** apresenta a localização cromossômica, posições inicial e final, além do identificador e do conjunto ao qual cada marcador pertence. Os conjuntos analisados incluem marcadores relacionados à cor dos olhos, genes envolvidos na pigmentação, marcadores de características capilares, SNPs do painel 34-plex, microhaplótipos utilizados para inferência de ancestralidade e marcadores associados à tonalidade da pele.

Tabela 3 - Coordenadas genômicas dos marcadores analisados. A tabela apresenta as informações de cromossomo, posição inicial e final, identificador do marcador (ID) e o conjunto ao qual pertence. Os conjuntos incluem marcadores associados à cor dos olhos, genes de pigmentação, características capilares, SNPs do painel 34-plex, microhaplótipos para inferência de ancestralidade e marcadores relacionados à tonalidade da pele.

Cromossomo	Posição inicial	Posição final	ID	Conjunto
chr8	3147516	3147517	rs11136584	Marcadores de olhos
chr15	28111712	28111713	rs1129038	
chr15	58508733	58508734	rs139083678	
chr2	221474113	221474114	rs1430211	
chr8	3339348	3339349	rs1444613	
chr3	151252897	151252898	rs1920395	
chr11	27456317	27456318	rs35177044	
chr5	33570495	33570496	rs396559	
chr4	24069421	24069422	rs4141480	
chr1	241609118	241609119	rs613032	
chr8	4976707	4976708	rs7007684	
chr15	100154980	100154981	rs78510225	
chr12	117554073	117554074	rs816188	
chr3	149131324	149131325	rs939249	
chr20	34194569	34269344	ASIP	Genes de pigmentação
chr9	16409503	16870843	BNC2	
chr8	2935353	4994972	CSMD1	
chr13	94436811	94479682	DCT	
chr16	89948697	89968060	DEF8	
chr15	28111040	28322179	HERC2	
chr6	391739	411443	IRF4	
chr12	88492793	88580851	KITLG	
chr1	235661041	235883724	LYST	
chr16	89912119	89920973	MC1R	
chr19	3538261	3574290	MFSD12	
chr3	69739456	69968336	MITF	
chr15	27754875	28099315	OCA2	
chr2	25160853	25168903	POMC	
chr14	92322581	92501481	SLC24A4	

chr15	48120990	48142672	SLC24A5	
chr5	33944623	33984693	SLC45A2	
chr11	69048932	69136316	TPCN2	
chr11	89177875	89295759	TYR	
chr9	12685439	12710285	TYRP1	
chr13	101950795	101950796	rs1033856	Marcadores de cabelo
chr2	195031893	195031894	rs12622325	
chr15	28165344	28165345	rs12916300	
chr1	101849657	101849658	rs1415105	
chr12	23956966	23956967	rs145940050	
chr5	33951587	33951588	rs16891982	
chr15	32854915	32854925	rs202139764	
chr15	48913058	48913059	rs34099850	
chr6	11184247	11184248	rs3734400	
chr2	237540363	237540364	rs3817362	
chr3	76526701	76526702	rs3913574	
chr16	89218910	89218911	rs4785768	
chr7	55127931	55127932	rs4947979	
chr11	64256583	64256584	rs588177	
chr16	3969528	3969529	rs61455471	
chr4	42894094	42894095	rs6815913	
chr3	76716585	76716586	rs699455	
chr3	122877676	122877677	rs7629139	
chr2	211805295	211805300	rs76536952	
chr5	34786730	34786731	rs7716361	
chr8	130908289	130908290	rs7843127	
chr1	40568446	40568447	rs3130316	SNPid 34-plex
chr1	169896172	169896173	rs1493232	
chr2	191195903	191195904	rs11142497	
chr2	21338453	21338454	rs907100	
chr3	185690684	185690685	rs3917	
chr3	57378604	57378605	rs2069949	
chr4	105249223	105249224	rs1528460	
chr4	52462237	52462238	rs10495407	
chr5	137611112	137611113	rs938283	
chr5	68957459	68957460	rs9637454	
chr6	26097011	26097012	rs170548	
chr6	76657658	76657659	rs6444724	
chr7	22392313	22392314	rs93834	
chr7	96527242	96527243	rs820869	
chr8	89751033	89751034	rs9818758	
chr8	19736762	19736763	rs2347915	
chr9	16053047	16053048	rs1360288	
chr9	133255426	133255427	rs2180715	
chr10	114376767	114376768	rs1354045	
chr10	45853876	45853877	rs1438277	

chr11	110086234	110086235	rs10461844	
chr11	62276438	62276439	rs861353	
chr12	88876551	88876552	rs10431792	
chr12	23140642	23140643	rs9409340	
chr13	103320596	103320597	rs154659	
chr13	32974647	32974648	rs735155	
chr14	63452538	63452539	rs1491087	
chr14	94378163	94378164	rs12326271	
chr15	47528244	47528245	rs7520386	
chr16	18237192	18237193	rs9876402	
chr17	76937134	76937135	rs7655302	
chr18	16112390	16112391	rs17108290	
chr19	45362243	45362244	rs870347	
chr20	61729438	61729439	rs887136	
chr1	11091432	11091663	mh01KK_212	Microhaplótipos para ancestralidade
chr2	217642124	217642401	mh02KK_215	
chr3	85410842	85411084	mh03KK_239	
chr4	71785543	71785722	mh04KK_238	
chr5	106256456	106256690	mh05KK_231	
chr6	160592654	160592901	mh06KK_236	
chr7	76484324	76484566	mh07KK_237	
chr8	6722374	6722635	mh08KK_244	
chr10	39182463	39182690	mh10KK_210	
chr11	109717984	109718219	mh11KK_249	
chr12	22503332	22503601	mh12KK_213	
chr2	237497722	237497723	rs10186010	Marcadores de pele
chr6	474404	474405	rs11242905	
chr10	13477513	13477514	rs12572522	
chr17	33835343	33835344	rs12602228	
chr15	41988568	41988569	rs141233250	
chr7	89298256	89298257	rs146113585	
chr5	33951587	33951588	rs16891982	
chr13	107694069	107694070	rs16970930	
chr11	88987149	88987150	rs187771782	
chr14	63907535	63907536	rs1953241	
chr4	17757332	17757333	rs2159309	
chr8	62099085	62099086	rs33937183	
chr20	53964866	53964867	rs376105	
chr8	14970595	14970596	rs73666954	
chr2	239162946	239162947	rs75061417	
chr15	45602839	45602840	rs76175998	
chr13	74086949	74086950	rs79748342	
chr5	124658815	124658816	rs9327337	
chr6	161890360	161890361	rs9355951	

3.2.4 Chamada de base, alinhamento e chamada de variantes

Os arquivos POD5 gerados no sequenciamento serão convertidos em arquivos BAM não alinhados utilizando o *Dorado basecaller* v0.9.0 (Nanoporetech, 2023a) com as opções de modelo HAC (*High Accuracy*) e *--modified-bases 5mCG_5hmCG* para a detecção de metilação e hidroximetilação em dinucleotídeos CpG. O controle de qualidade do sequenciamento será realizado com o *cramino* v0.15.0 (De Coster & Rademakers, 2023). Para o alinhamento dos arquivos BAM, será utilizado o *Dorado aligner*, que emprega o *minimap2* (Li, 2018) como base, tendo como genoma de referência o GRCh38 (GCA_000001405.15_GRCh38). O controle de qualidade e processamento dos arquivos BAM alinhados será realizado com o *Samtools* v1.21 (Li et al., 2009). As chamadas de variantes serão conduzidas com o *freebayes* v1.3.8 (Garrison & Marth, 2012) e o *Clair3* v1.0.10 (Zheng et al., 2022). Como o Clair3 não possui funcionalidade para chamadas de variantes em múltiplas amostras (ao contrário do *freebayes*), a opção *--gvcf* será utilizada para gerar arquivos no formato gvcf, os quais serão processados para chamadas de variantes comuns com o *GLnexus* v1.4.1 (M. F. Lin et al., 2018; Yun et al., 2021), usando configurações específicas para gvcf gerados pelo *Clair3*. Os arquivos BAM também serão submetidos à haplotipagem utilizando o *Longphase* (J.-H. Lin et al., 2022; Zheng et al., 2022).

3.2.5 Determinação de haplótipos longos

A determinação de haplótipos estendidos envolvendo genes de pigmentação completos foi realizada por meio de montagem *de novo* com correção de fase utilizando PECAT (Nie et al., 2024). Resumidamente, *samtools fastq* foi utilizado para extrair as leituras alinhadas para cada gene nos arquivos BAM de cada amostra. Estes arquivos então foram utilizados como entrada para PECAT que será executado utilizando as opções de faseamento com *clair3* e polimento com *medaka* v1.7.2 (Nanoporetech, 2023b). Os haplótipos montados foram alinhados ao genoma de referência utilizando *minimap2* com os seguintes parâmetros: '*--ax asm20 --B 2 --E 3,1 --O 6,100 --cs --K 5G* e os as posições de início e final de cada *contig* foram determinadas utilizando *bedtools* v2.31 (Quinlan & Hall, 2010).

3.2.6 Análise de metilação

Os arquivos *bedMethyl* de metilação do genoma inteiro, resolvidos por haplótipo, foram gerados com o *modkit* v0.4.2 (Nanoporetech, 2023c) usando o programa *pileup* com os sinalizadores *–cpg*, *--ref*, *--ignore h*, *--combine-strands*, e *--partition-tag HP*. As fitas foram combinadas para um valor de contagem de leitura usando a opção *--combine_strands* para cada CpG correspondente aos CpGs na referência GRCh38. Os arquivos bedMethyl de metilação do genoma inteiro, resolvidos por haplótipos, foram gerados com o modkit utilizando o programa pileup com os sinalizadores *--cpg, --ref, --partition-tag HP*, *--combine-strands* e *--ignore h*. As fitas foram combinadas para um valor de contagem de leitura usando a opção --combine-strands

para cada CpG correspondente aos CpGs na referência GRCh38. A partir dos arquivos bedMethyl resolvidos por haplótipos, a análise de metilação diferencial entre haplótipos foi realizada com o programa *modkit dmr pair*. Para cada indivíduo, os arquivos de metilação específicos de cada haplótipo foram comparados utilizando os sinalizadores *--ref,* e *--base C*. A análise foi realizada considerando apenas regiões CpG e usando a referência GRCh38. A entropia da metilação foi calculada utilizando *modkit entropy*. A análise foi conduzida com os sinalizadores *--cpg, --ref,* e *--regions*, considerando a referência GRCh38. Os valores de entropia foram calculados para cada gene e indivíduo, permitindo a avaliação da variabilidade nos padrões de metilação ao longo do genoma.

4. RESULTADOS E DISCUSSÃO

4.1 Amostras coletadas

4.1.1 Características fenotípicas das amostras

Até o momento, foram coletadas amostras de 14 pacientes, abrangendo diferentes tipos de materiais biológicos. Para cada participante, foram obtidas amostras de pele (n=14), sangue (n=12) e saliva (n=12). Todas as participantes são do sexo feminino, sendo 13 destras e 1 canhota. A seguir, apresentamos gráficos que ilustram as principais características fenotípicas e étnicas das amostras coletadas.

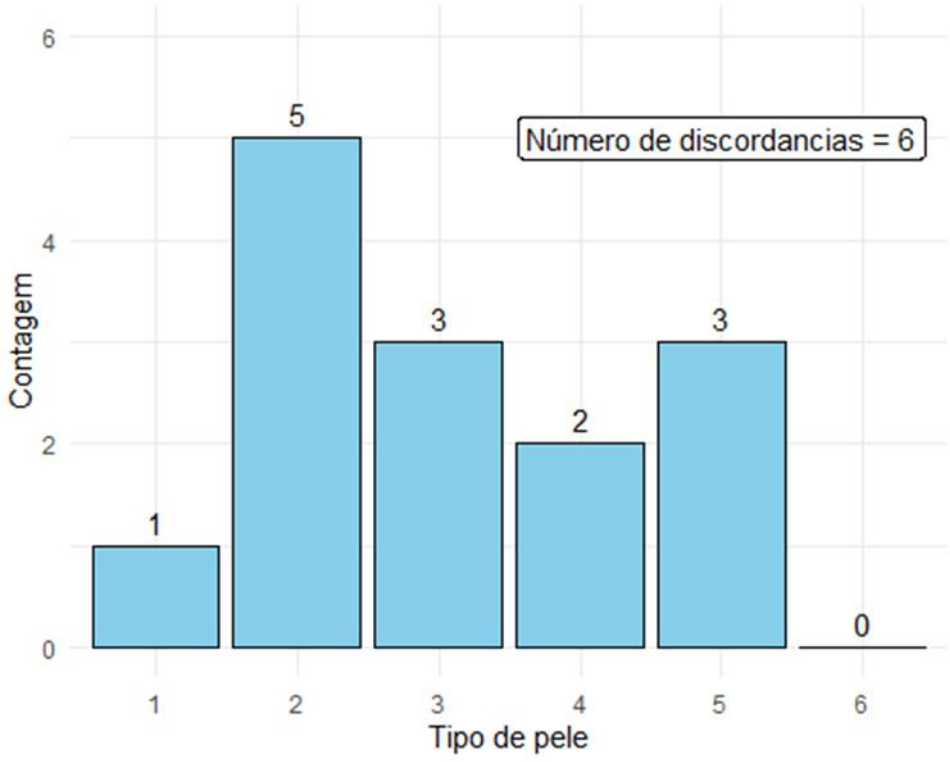


Figura 4 - Gráfico de barras dos tipos de pele coletados.

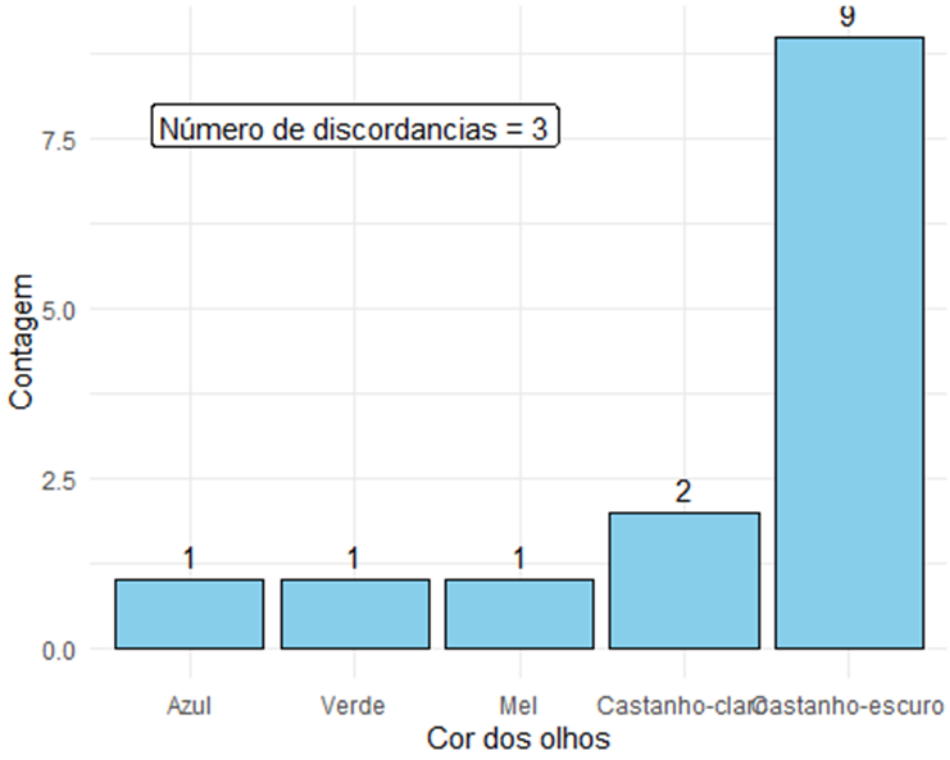


Figura 5 - Gráfico de barras da cor dos olhos das amostras coletadas.

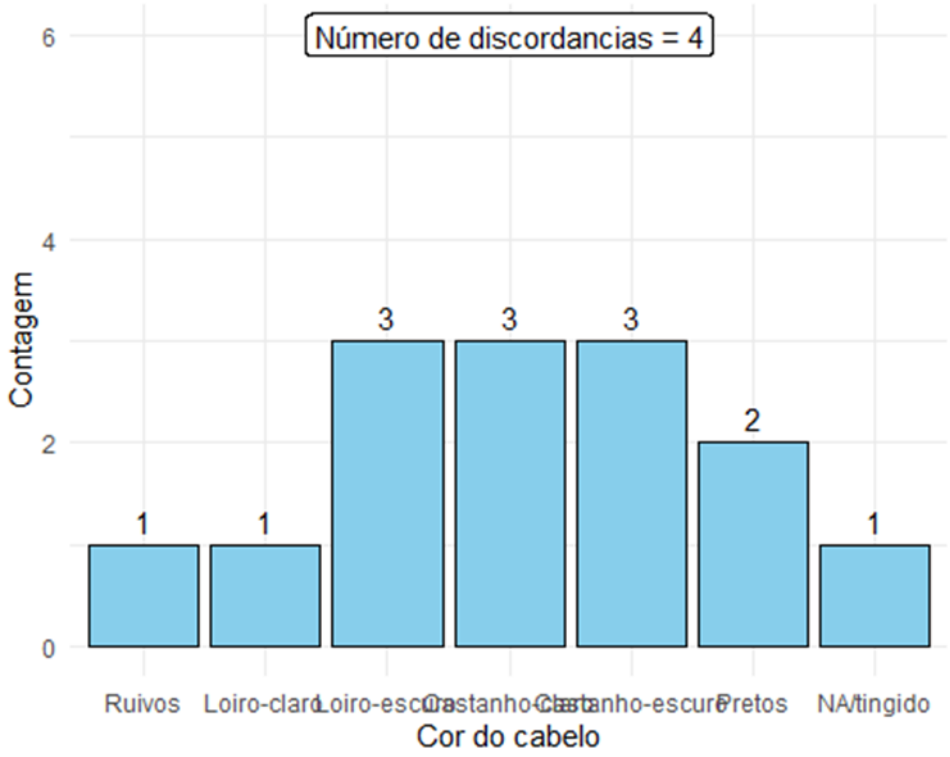


Figura 6 - Gráfico de barras das cores de cabelo das amostras coletadas.

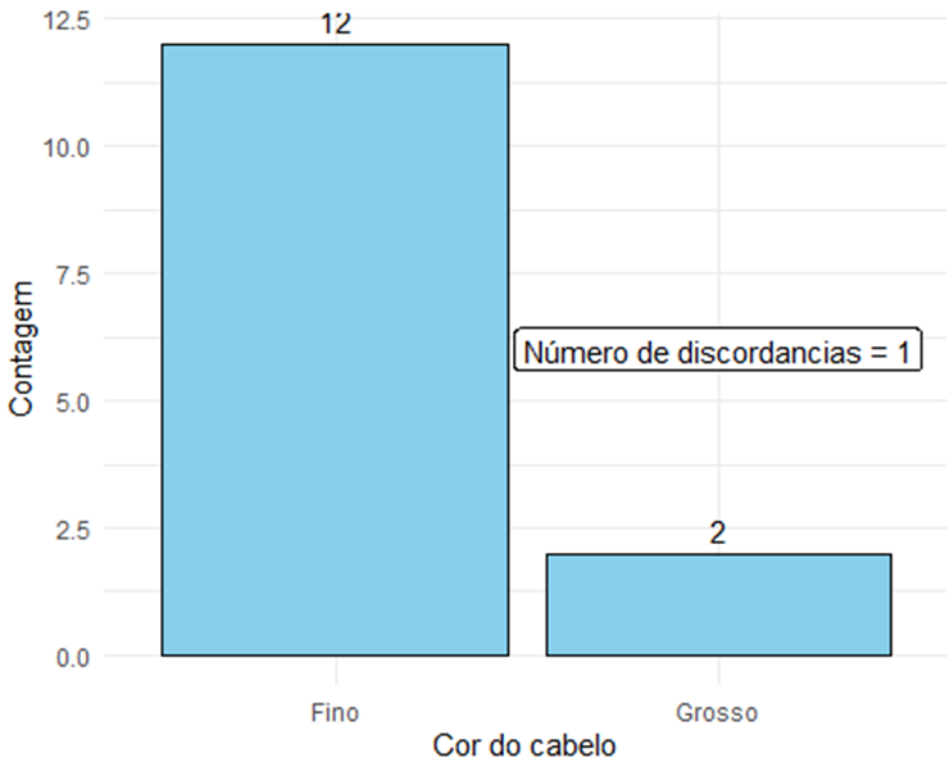


Figura 7 - Gráfico de barras dos tipos de cabelo das amostras coletadas.

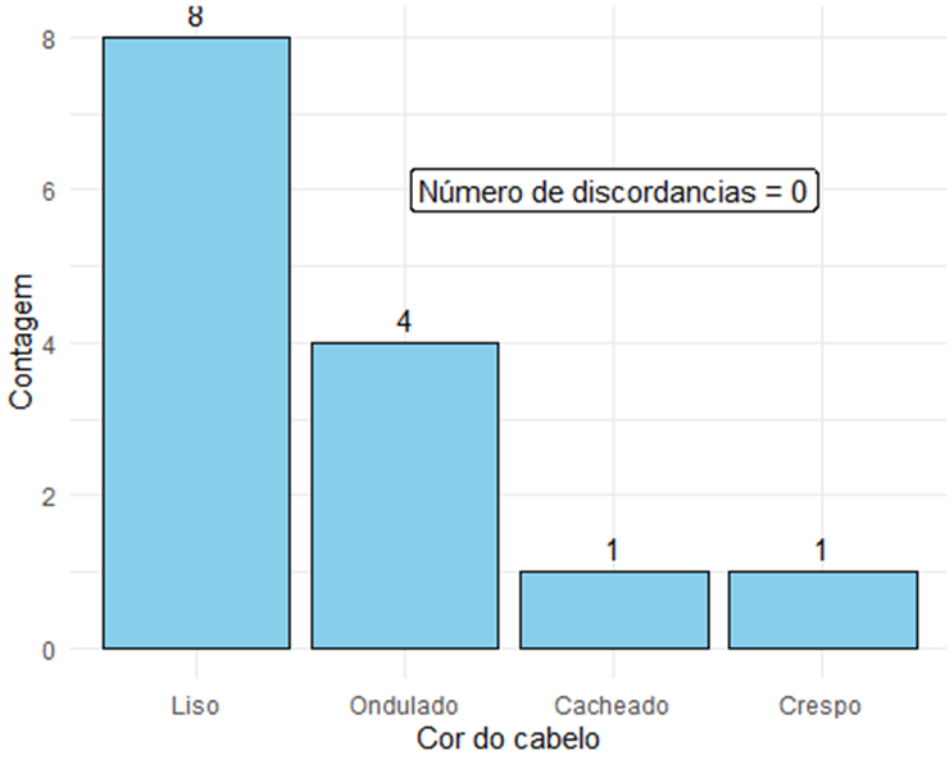


Figura 8 - Gráfico de barras dos tipos de cabelo das amostras coletadas.

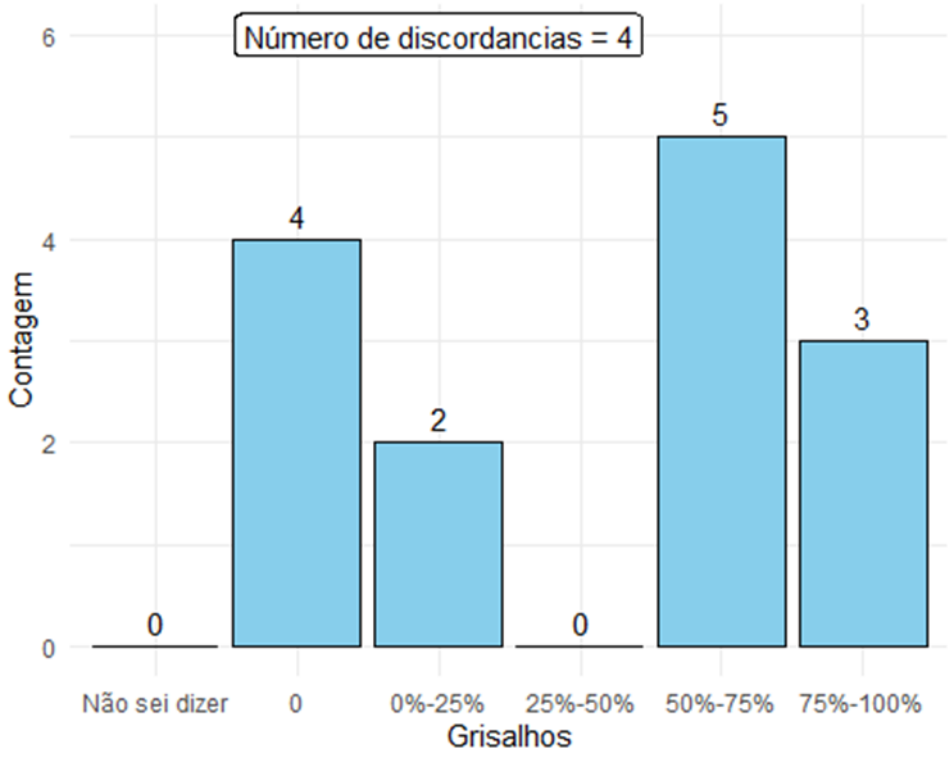


Figura 9 - Gráfico de barras da porcentagem de grisalhos nas amostras coletadas.

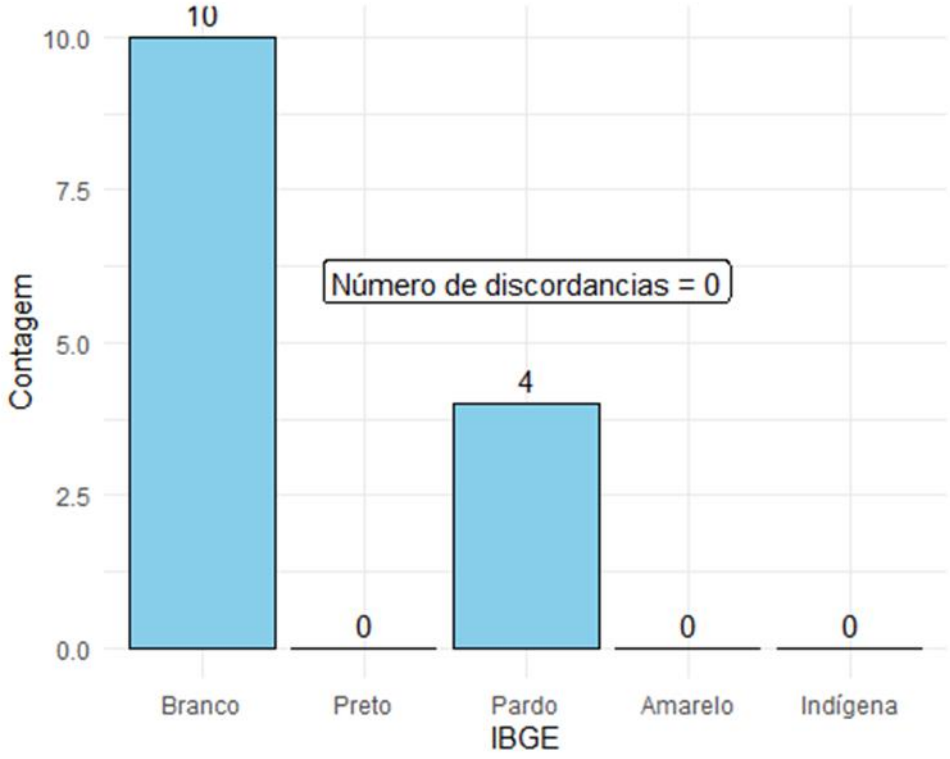


Figura 10 - Gráfico de barras da classificação do IBGE das amostras coletadas.

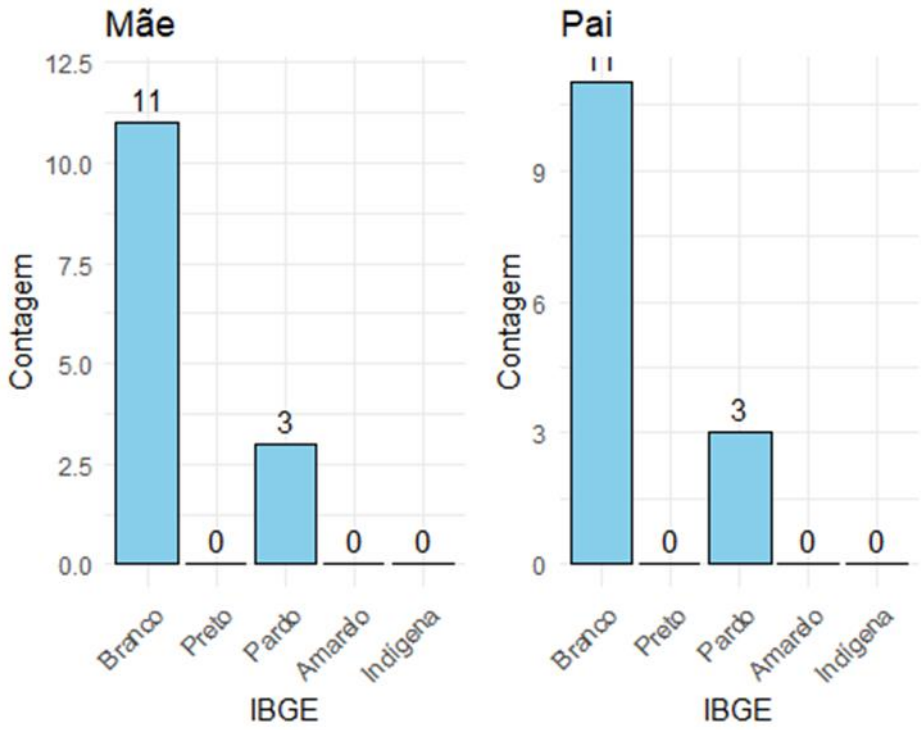


Figura 11 - Gráfico de barras da classificação do IBGE dos pais das amostras coletadas.

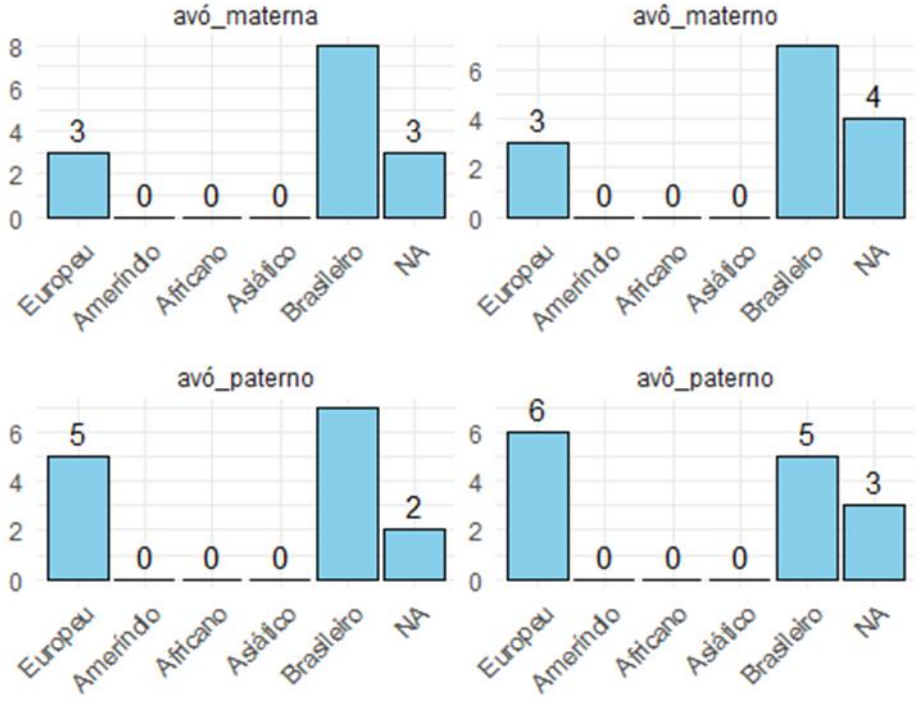


Figura 12 - Gráfico de barras das origens dos avós das amostras coletadas.

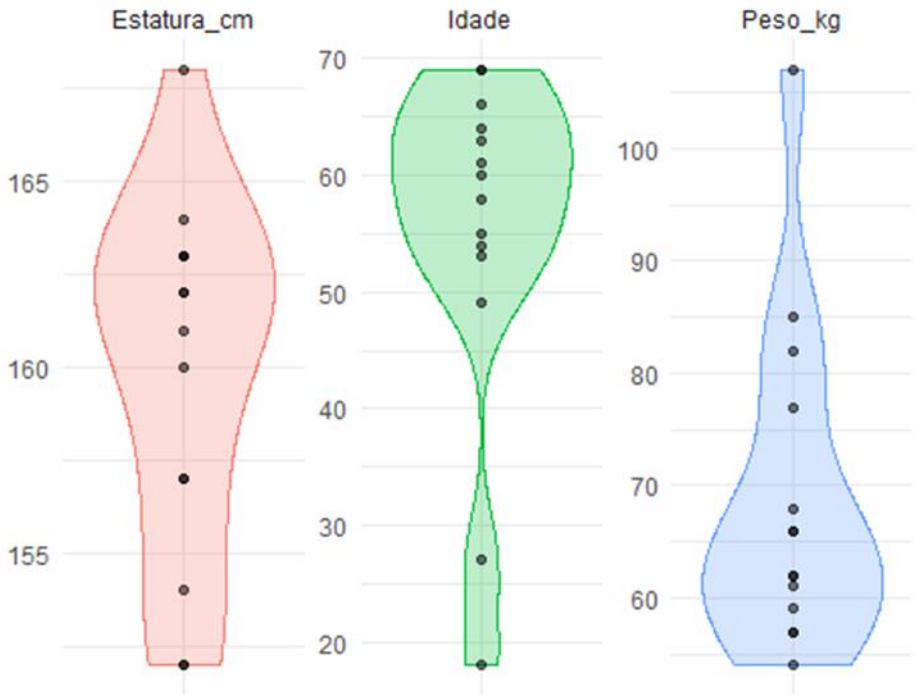


Figura 13 – Gráfico violino da estatura (cm), idade (ano) e peso (kg) das amostras coletadas.

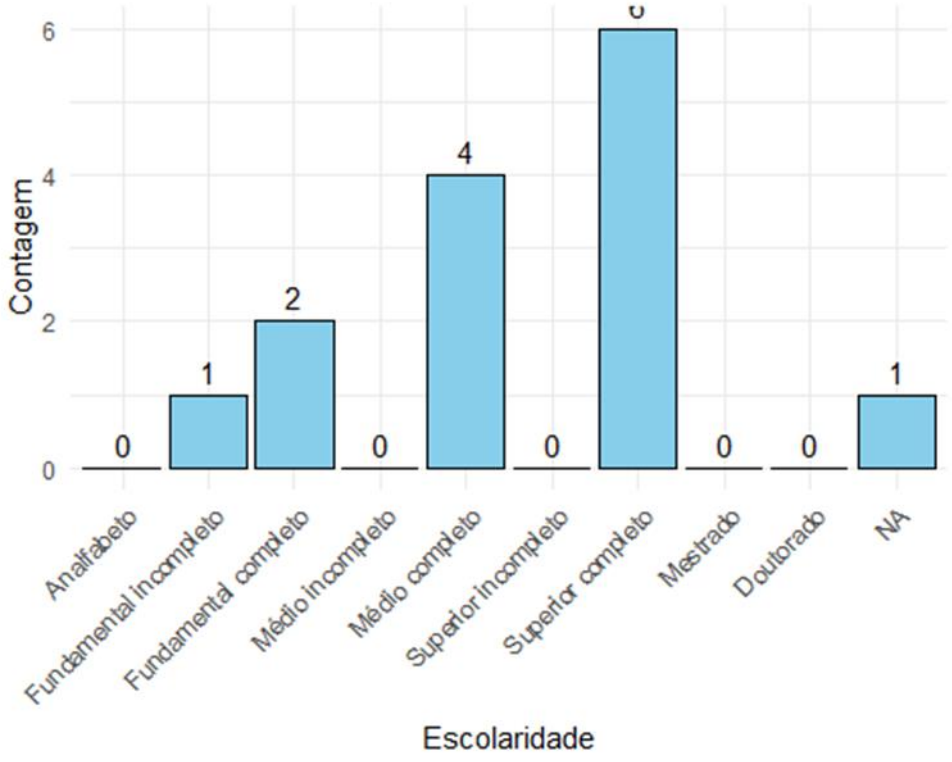


Figura 14 - Gráfico de barras do grau de escolaridade das amostras coletadas.

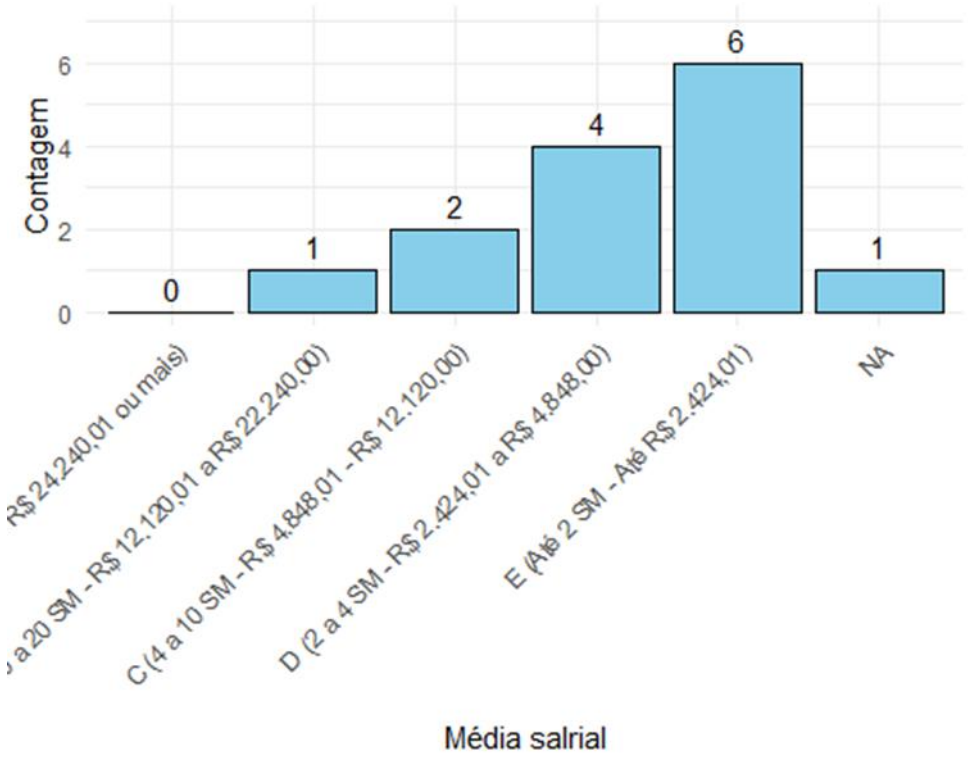


Figura 15 -Gráfico de barras da média salarial das amostras coletadas.

4.1.2 Quantificação das amostras

A quantificação das amostras de DNA foi realizada utilizando os métodos Qubit BR assay e Nanodrop, permitindo a avaliação da concentração e da pureza do DNA extraído de diferentes tipos de tecido, incluindo sangue, saliva e pele. Os resultados revelaram variações significativas entre as amostras, tanto em termos de concentração absoluta quanto nos índices espectrofotométricos A260/280 e A260/230.

As amostras de sangue apresentaram, em geral, concentrações elevadas de DNA, com algumas atingindo o limite superior de detecção do Qubit (1000 ng/µL), como observado para as amostras 602, 604 e 607. No entanto, algumas amostras exibiram discrepâncias entre as quantificações pelo Qubit e pelo Nanodrop, o que pode indicar a presença de contaminantes que influenciam a leitura espectrofotométrica. Por exemplo, a amostra 601-Blood apresentou 800 ng/µL pelo Qubit e 1215,6 ng/µL pelo Nanodrop, sugerindo possível contaminação por proteínas ou outros compostos. Além disso, algumas amostras apresentaram índices A260/280 e A260/230 fora dos valores esperados para DNA puro, como 611-Blood, que mostrou A260/280 de 1,74 e A260/230 de 1,13, indicando possível contaminação por proteínas e sais.

As amostras de saliva exibiram uma variabilidade maior na quantificação de DNA, com algumas amostras apresentando concentrações muito baixas, como 607-Saliva e 608-Saliva, que ficaram abaixo do limite de detecção do Qubit (0 ng/µL). Outras amostras, como 604-Saliva e

605-Saliva, apresentaram valores significativamente diferentes entre os dois métodos, indicando possíveis interferências na leitura espectrofotométrica. O índice A260/280 também variou consideravelmente, sendo que algumas amostras apresentaram valores abaixo do esperado, como 606-Saliva (1,5), sugerindo contaminação por proteínas.

Nas amostras de pele, foram observadas altas concentrações de DNA em diversas amostras, com várias atingindo o limite superior de detecção do Qubit, como 607-Skin, 608-Skin e 609-Skin (1000 ng/μL). O índice de pureza foi relativamente mais estável em comparação às amostras de saliva, com a maioria das amostras apresentando A260/280 entre 1,8 e 2,2 e A260/230 acima de 1,8, sugerindo menor contaminação.

Por fim, algumas amostras apresentaram problemas durante a coleta, resultando em ausência de dados na quantificação, como 603-Blood, 610-Blood e 611-Saliva. Aos todos estes resultados indicam a necessidade de otimização na coleta e extração do DNA, além da importância de utilizar métodos complementares para avaliação da qualidade das amostras antes de análises subsequentes.

Tabela 4 - Tabelas de quantificação das amostras utilizando Qubit BR assay e Nanodrop. As linhas cinzas representam amostras com problemas durante as coletas, linhas vermelhas representam amostras com valores de A260/280 e A260/230 indicando baixa pureza no DNA. Amostras que apresentaram quantificações acima ou abaixo dos limites de detecção do Qubit foram marcadas em negrito e representam como 1000 ng/uL (Para concentrações acima) e 0 ng/uL (Para concentrações abaixo).

Amostra	Tecido	Volume (uL)	Qubit BR (ng/uL)	Nanodrop (ng/uL)	A260/280	A260/230	Massa total Qubit (ng)	Massa total Nanodrop (ng)
600	Blood	500	630.000	125.200	1.43	1.68	315000.00	62600.00
600	Saliva	500	6.680	10.880	2.07	1.74	3340.00	5440.00
600	Skin	500	99.700	165.500	2.19	1.92	49850.00	82750.00
601	Blood	150	800.000	1215.600	2.39	1.86	120000.00	182340.00
601	Saliva	150	25.300	48.600	2.62	1.83	3795.00	7290.00
601	Skin	150	750.000	588.000	2.32	1.85	112500.00	88200.00
602	Blood	150	**1000.000**	737.800	2.34	1.91	**150000.00**	110670.00
602	Saliva	150	127.000	124.400	1.95	1.91	19050.00	18660.00
602	Skin	150	500.000	539.300	2.21	1.88	75000.00	80895.00
603	Blood	150						
603	Saliva	150	35.800	66.490	1.92	1.89	5370.00	9973.50
603	Skin	150	160.000	246.700	2.17	1.93	24000.00	37005.00
604	Blood	150	**1000.000**				**150000.00**	0.00
604	Saliva	150	468.000	863.500	1.86	1.99	70200.00	129525.00
604	Skin	150	1000.000	1735.600	1.9	2.32	150000.00	260340.00
605	Blood	150	560.000	346.600	1.9	2.29	84000.00	51990.00
605	Saliva	150	217.000	81.000	1.85	2.17	32550.00	12150.00
605	Skin	150	980.000	1274.900	1.89	2.22	147000.00	191235.00
606	Blood	300	4.930	120.200	1.92	3.62	1479.00	36060.00

606	Saliva	300	7.330	22.400	1.5	0.92	2199.00	6720.00
606	Skin	150	850.000	681.400	1.9	2.23	127500.00	102210.00
607	Blood	300	830.000	1782.200	1.89	2.3	249000.00	534660.00
607	Saliva	300	**0.000**	0.600	1.1	0.23	**0.00**	180.00
607	Skin	150	**1000.000**	2117.300	1.89	2.23	**150000.00**	317595.00
608	Blood	300	580.000	78.800	1.8	1.39	174000.00	23640.00
608	Saliva	300	**0.000**	0.700	0.81	0.33	**0.00**	210.00
608	Skin	150	900.000	1071.900	1.88	2.22	135000.00	160785.00
609	Blood	300	420.000	671.000	1.9	1.79	126000.00	201300.00
609	Saliva	300	93.900	69.300	1.73	0.86	28170.00	20790.00
609	Skin	150	**1000.000**	2066.400	1.86	2.14	**150000.00**	309960.00
610	Blood	300						
610	Saliva	300	118.000	13.100	1.43	0.34	35400.00	3930.00
610	Skin	150	740.000	1390.400	1.88	2.24	111000.00	208560.00
611	Blood	300	550.000	1309.200	1.74	1.13	165000.00	392760.00
611	Saliva	300						
611	Skin	150	950.000	1237.900	1.86	2.06	142500.00	185685.00
612	Blood	300	453.000	387.800	1.91	1.67	135900.00	116340.00
612	Saliva	300	**0.000**	404.200	1.89	2	0.00	121260.00
612	Skin	150	500.000	10.800	1.44	0.93	75000.00	1620.00
613	Blood	300	36.700	5.500	1.42	-0.41	11010.00	1650.00
613	Saliva	300	2.940	5.900	2.02	-2	882.00	1770.00
613	Skin	150	540.000	564.200	1.78	1.72	81000.00	84630.00

4.2 Estabelecimento dos procedimentos computacionais com a amostra HG002

Com o objetivo de padronizar os procedimentos analíticos empregados neste estudo, utilizamos a amostra HG002 como referência para o estabelecimento e validação do pipeline computacional. Essa amostra foi selecionada por sua alta qualidade e ampla caracterização no contexto do consórcio *Genome in a Bottle* (GIAB), servindo como base confiável para testes metodológicos. As etapas analisadas incluem o processo de chamada de base (basecalling), a detecção de variantes genômicas (variant calling) e a análise de metilação de DNA. Os resultados apresentados a seguir descrevem a aplicação dessas etapas, bem como os parâmetros utilizados e as ferramentas selecionadas, fornecendo uma estrutura sólida para a análise das demais amostras.

4.2.1 Pipeline para análise de dados de sequenciamento via Oxford Nanopore Technologies

O sequenciamento via ONT tem como produto arquivos que contêm a informação do sinal elétrico ao longo de todo experimento armazenado. Estes dados são classicamente gravados em formatos especiais chamados FAST5 ou POD5 (Formato mais moderno) e devem ser convertidos em arquivo FASTQ por meio do processo de chamada de bases (basecalling). ONT e a comunidade disponibilizam algumas opções de software para este processo sendo dorado v0.5.3

(https://github.com/nanoporetech/dorado) considerado o estado-da-arte para tal. Dentre as diferentes funcionalidades de dorado estão a capacidade de lidar com dados simplex ou duplex, trimagem de adaptadores, chamada de bases modificadas (análises epigenéticas), dentre outras. Embora seja possível utilizar arquivos FAST5, é recomendado a conversão para POD5 utilizando a biblioteca em python pod5 (https://pod5-file-format.readthedocs.io/en/latest/). Dorado permite que a saída seja arquivos FASTQ (formato de sequências) ou então BAM (formato de alinhamento), porém uma vez que estes BAM não estão alinhados é necessário realizar o alinhamento a um genoma de referência. O software permite o alinhamento utilizando minimap2, sendo esta etapa passível de ser realizada fora do dorado. Por fim deve-se utilizar samtools para ordenar e indexar este BAM. A **Figura 16** - Fluxograma de procedimentos para análise de dados de sequenciamento Oxford Nanopore.

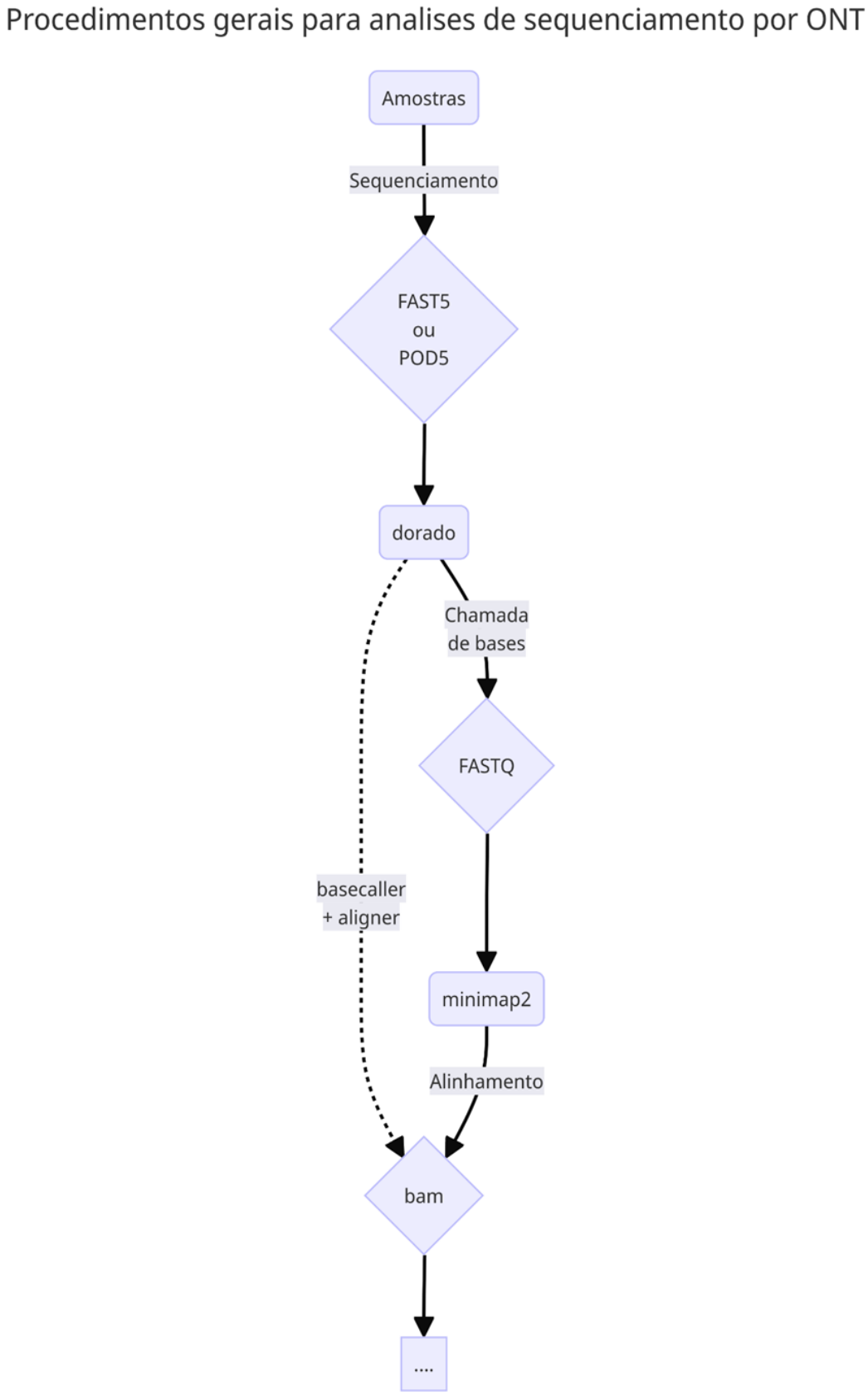


Figura 16 - Fluxograma de procedimentos para análise de dados de sequenciamento Oxford Nanopore.

4.2.2 Pipeline para chamada de variantes

Para chamadas de variantes utilizamos duas ferramentas: *freebayes* (https://github.com/freebayes/freebayes) e *Clair3* (https://github.com/HKU-BAL/Clair3). O *freebayes* é uma ferramenta popular de detecção de variantes genéticas de código aberto usada principalmente para identificar variantes de nucleotídeo único (SNVs) e pequenas inserções/deleções (indels) de dados de sequenciamento, especialmente de experimentos de sequenciamento de genoma completo e exoma completo. Ele foi projetado para trabalhar com genomas diploides e poliploides e pode lidar com conjuntos de dados pequenos e grandes com eficiência. *Freebayes* apresenta a possibilidade de lidar com múltiplas amostras ao mesmo tempo (comumente conhecido pelo termo de *joint caller*). Em geral, o *freebayes* é uma ferramenta boa e robusta para chamadas de SNP. No entanto, ela exige muita filtragem posterior pela qualidade (Coluna QUAL) etc. para obter bons conjuntos de SNP. Ou seja, ela se concentra na sensibilidade, não na especificidade, portanto, produz muitas variantes potenciais.

Já *Clair3* é uma ferramenta especializada projetada para identificar pequenas variantes da linha germinativa a partir de dados de sequenciamento de leitura longa. Ela utiliza uma abordagem híbrida que combina dois métodos principais para obter uma identificação precisa de variantes:

Chamada de empilhamento (***Pileup calling***): Esse método é eficiente para lidar rapidamente com a maioria dos candidatos a variantes. Ele provavelmente aproveita as informações de empilhamento, que representam as leituras alinhadas em cada posição genômica, para identificar variantes com eficiência, especialmente em regiões com cobertura suficiente e onde a identificação de variantes é relativamente simples.

Alinhamento completo (***Full-alignment***): Para candidatos a variantes mais complexas que não podem ser identificados com segurança usando apenas a abordagem de pileup, *Clair3* emprega uma estratégia de alinhamento completo. Isso envolve o alinhamento de leituras de forma mais abrangente ao genoma de referência, permitindo a caracterização precisa de variantes em regiões desafiadoras, o que pode melhorar a precisão e a recuperação.

Clair3 vem sendo amplamente utilizada em série de trabalhos publicados e inclusive sendo integrado aos *workflows* fornecidos pela própria Oxford Nanopore na plataforma EPI2ME (https://github.com/epi2me-labs/wf-human-variation). Entretanto *Clair3* não possui a funcionalidade de '*joint caller*' sendo necessário que a genotipagem ocorra em formato gVCF e depois a utilização do software *GLnexus* (https://github.com/dnanexus-rnd/GLnexus) para unificar em um VCF final. A figura a seguir contém um fluxograma das ferramentas escolhidas para realização desta etapa (**Figura 17**)

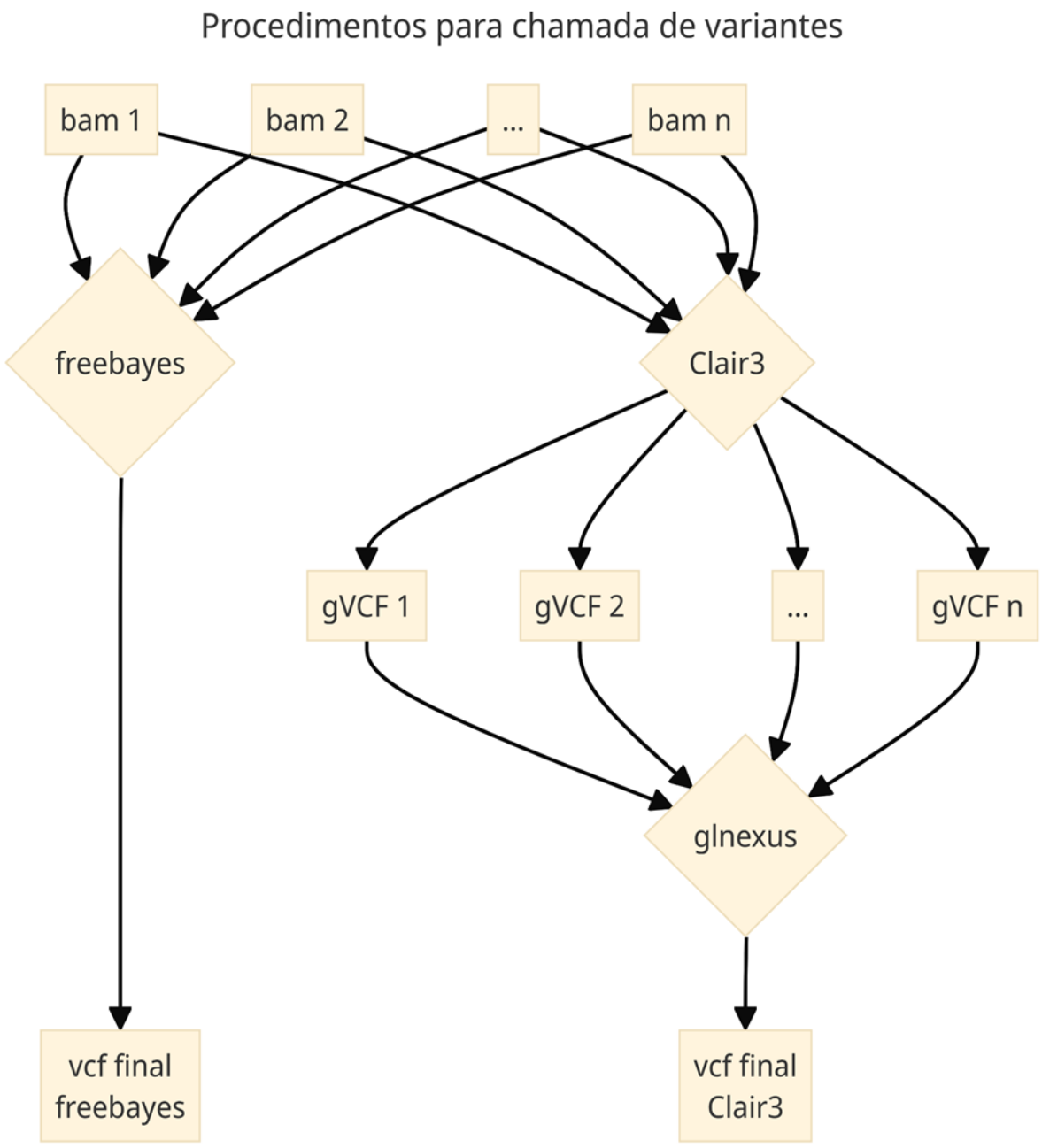


Figura 17 - Fluxograma de procedimentos para chamada de variantes em dados de sequenciamento Oxford Nanopore via freebayes e Clair3.

Para testar estas ferramentas utilizamos a amostra HG002 do GIAB e realizamos o mesmo procedimento para freebayes. Já para Clair3 o procedimento foi levemente alterado para o formato de saída ser VCF (sem necessidade de usar GLnexus). Esta abordagem apresenta o viés de ser para uma amostra única, devendo ser realizada para as 3 amostras do GIAB para avaliar a chamada em conjunto. A figura abaixo (**Figura 18**) mostra a cobertura do sequenciamento para HG002 mostrando valores acima de 20x para todos os genes, com a maior dos valores acima de 30x.

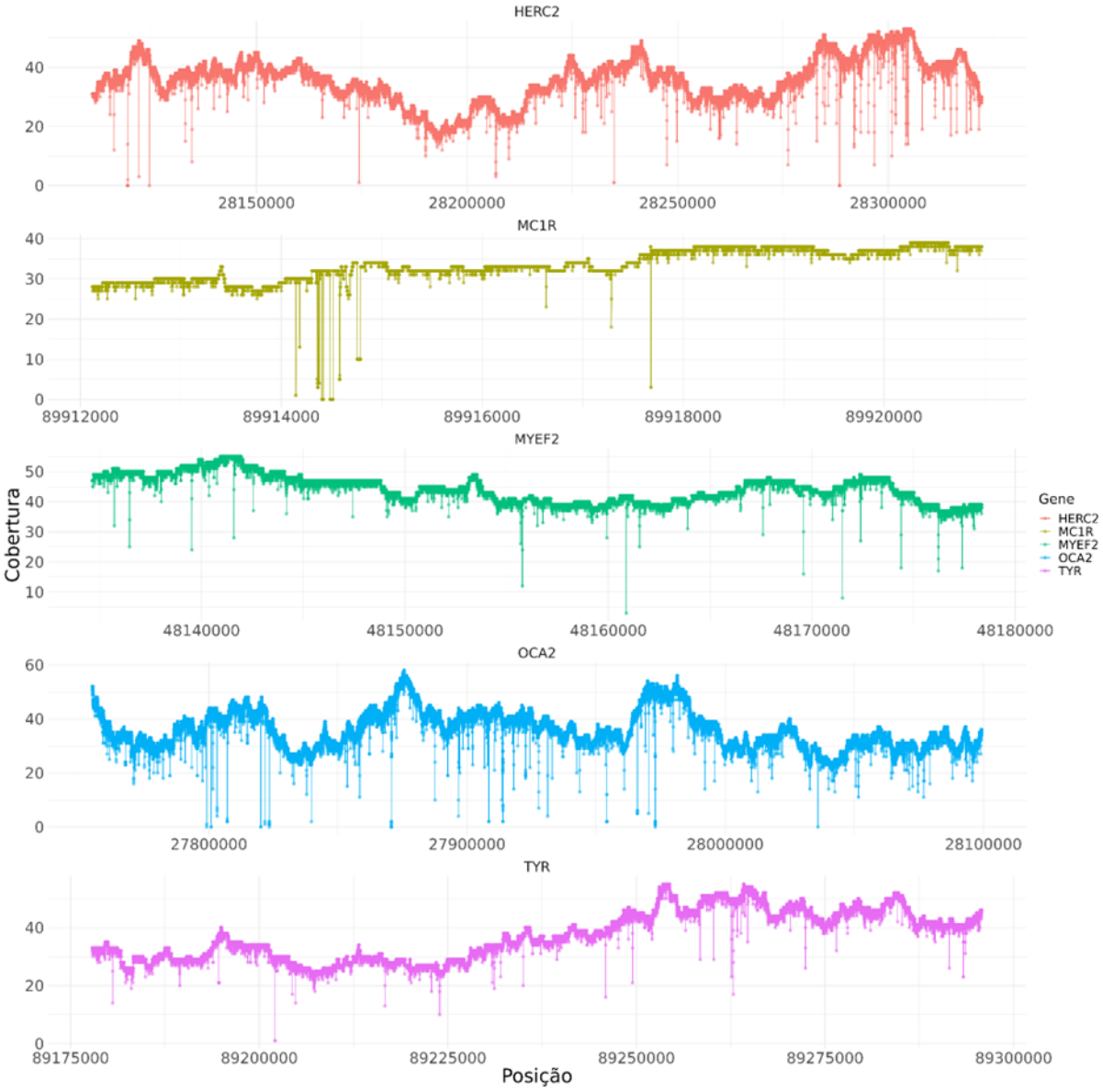


Figura 18 - Visualização da cobertura por posição dos genes para a amostra HG002 do GIAB.

A figura a seguir (**Figura 19**) mostra uma região para o gene TYR com cobertura superior a 50x. Nesta é possível notarmos que as variantes obtidas por ambos os softwares apresentam boa semelhança com as variantes observadas nas leituras. Entretanto, nesta imagem podemos ver que freebayes apresenta maior número de falsos positivos que clair3.

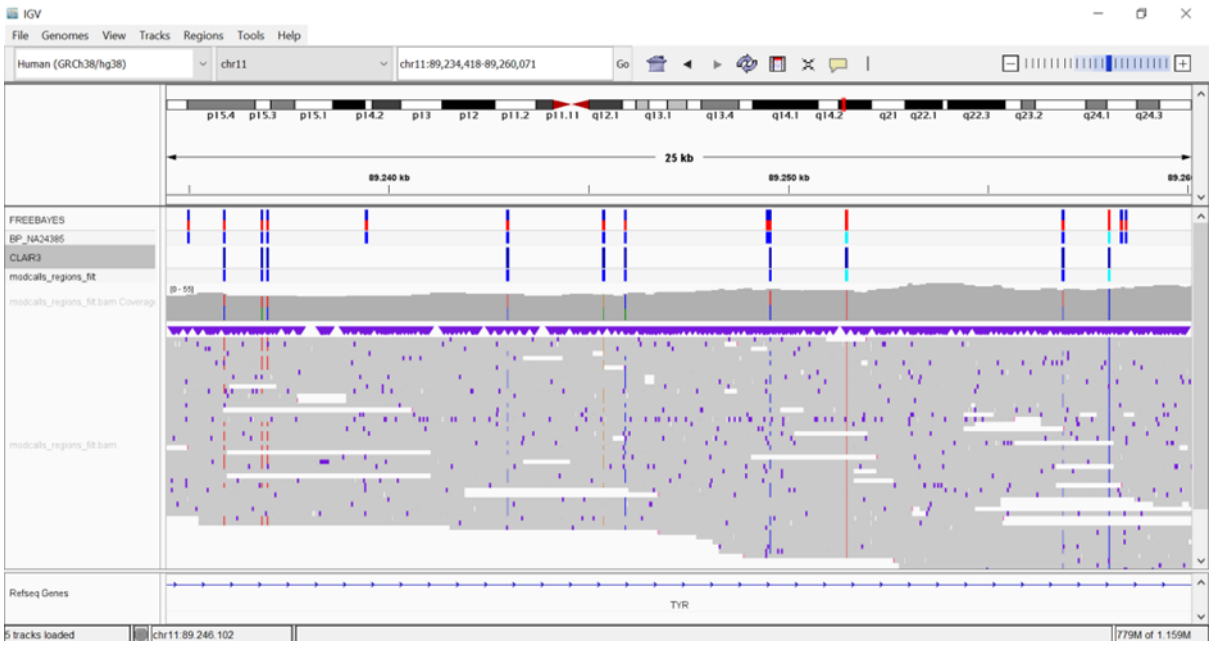


Figura 19 - Visualização da região do gene TYR no IGV. As trilhas superiores contêm as variantes para Freebayes e Clair3.

Para avaliar a acurácia do método baixamos o VCF fornecido pelo GIAB para amostra HG002 (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/latest/GRCh38/) e testamos os resultados obtidos por ambos os métodos considerando este VCF como o correto. Este VCF contém 984 variantes anotadas na região dos genes por nós analisados, enquanto os VCFs de freebayes e clair3 contém 1283 e 1436, respectivamente. As figuras c contém as comparações para os genes avaliados.

Figura 20 - Visualização da região do gene MC1R no IGV para a comparação dos VCFs com o fornecido por GIAB.

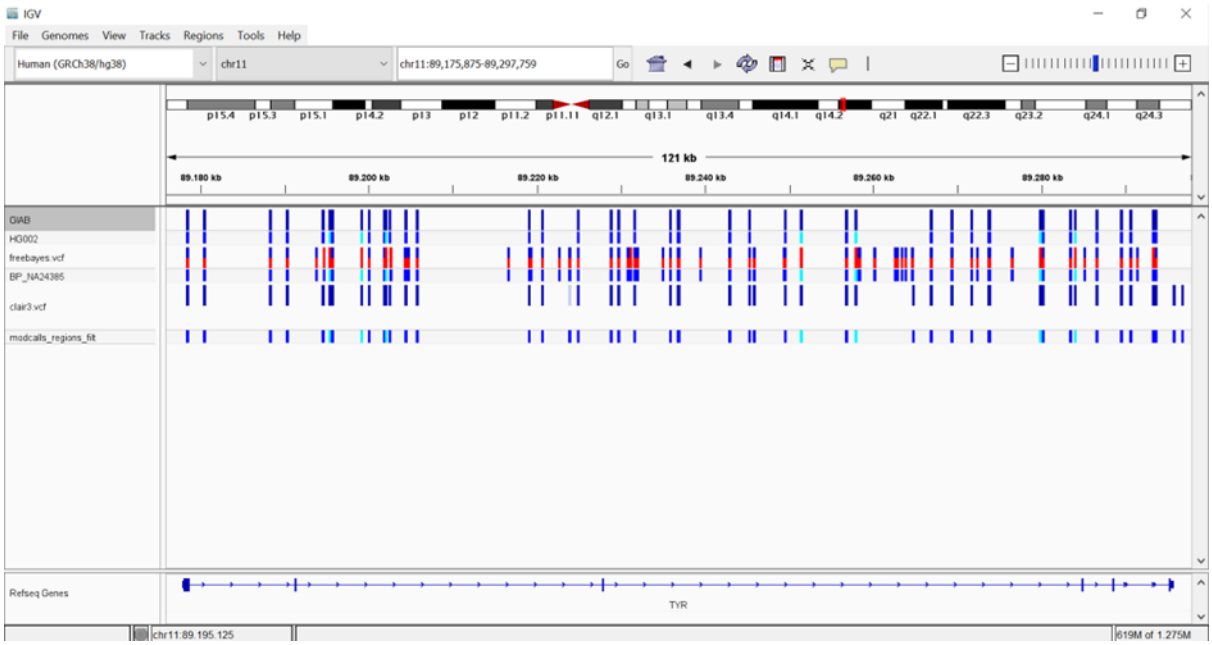


Figura 21 – Visualização da região do gene TYR no IGV para a comparação dos VCFs com o fornecido por GIAB.

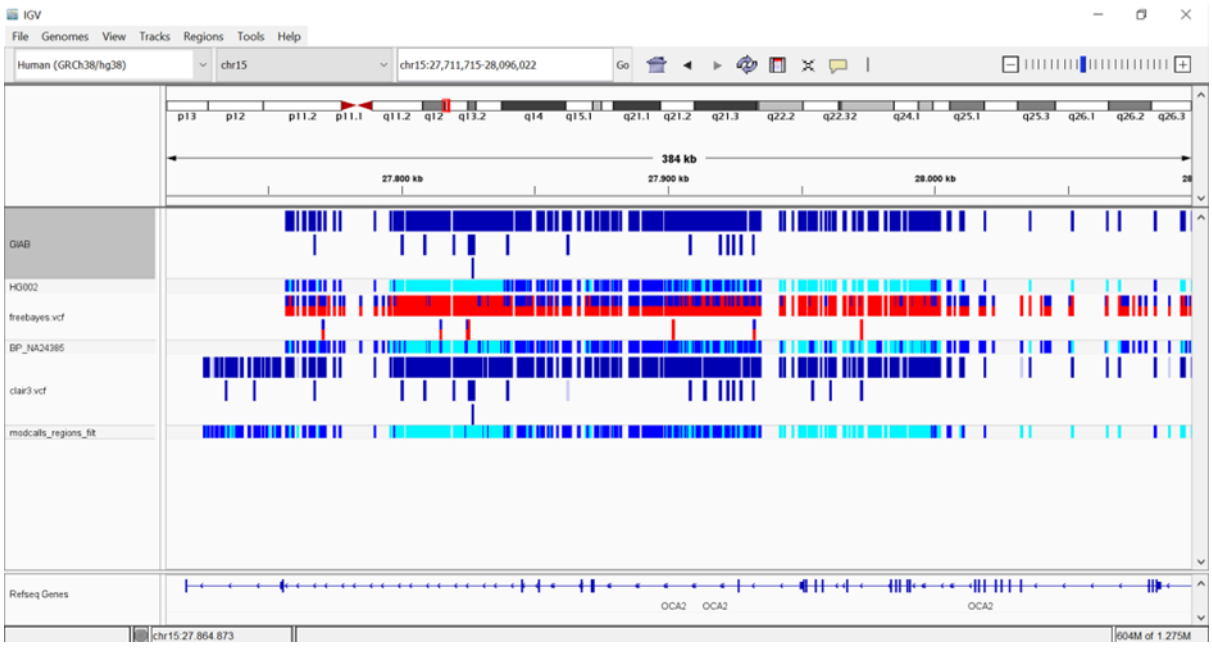


Figura 22 - Visualização da região do gene OCA2 no IGV para a comparação dos VCFs com o fornecido por GIAB.

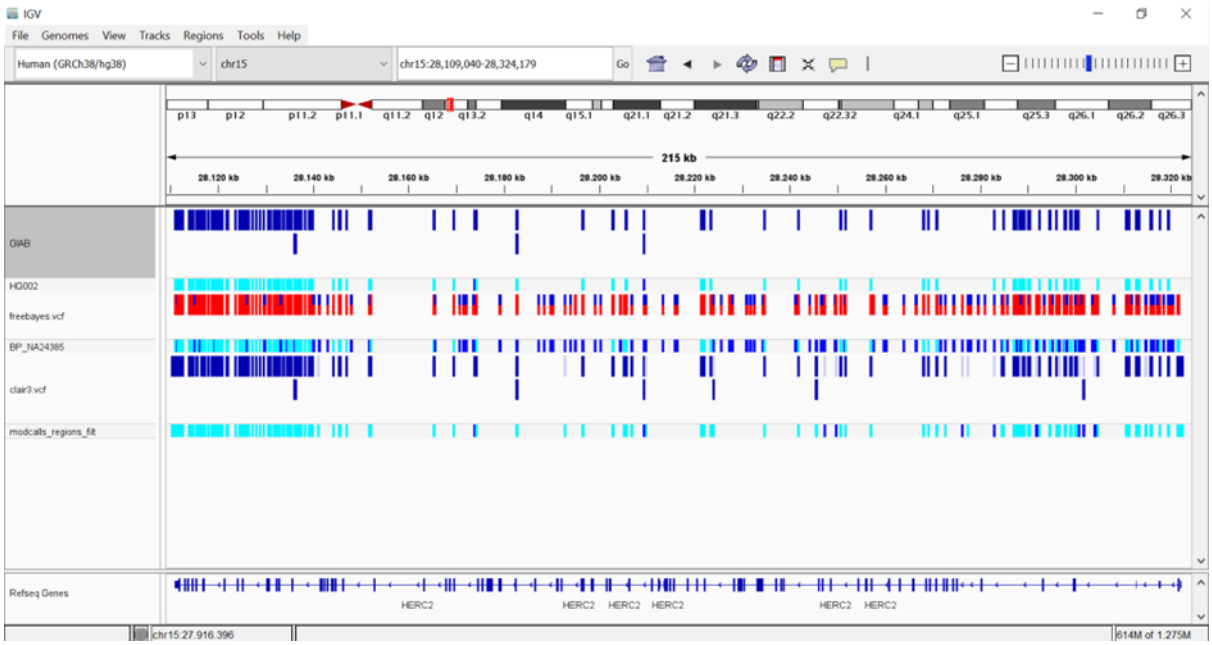


Figura 23 - Visualização da região do gene HERC2 no IGV para a comparação dos VCFs com o fornecido por GIAB.

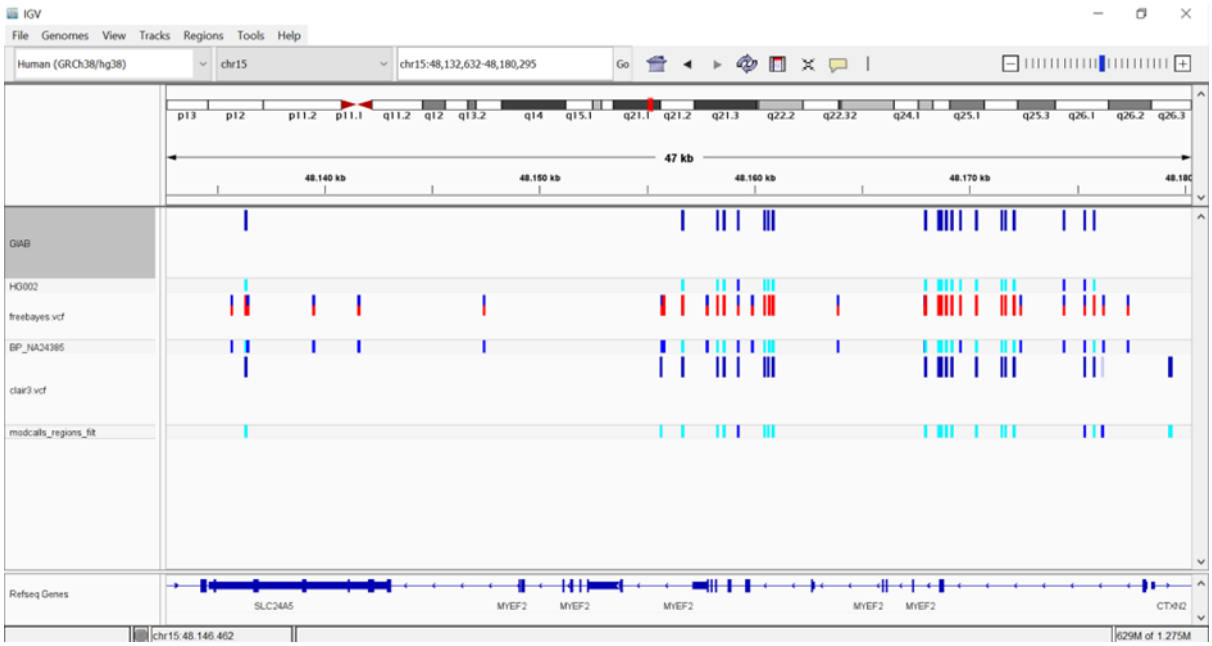


Figura 24 - Visualização da região do gene MYEF2 no IGV para a comparação dos VCFs com o fornecido por GIAB.

4.2.3 Procedimentos para análise da metilação

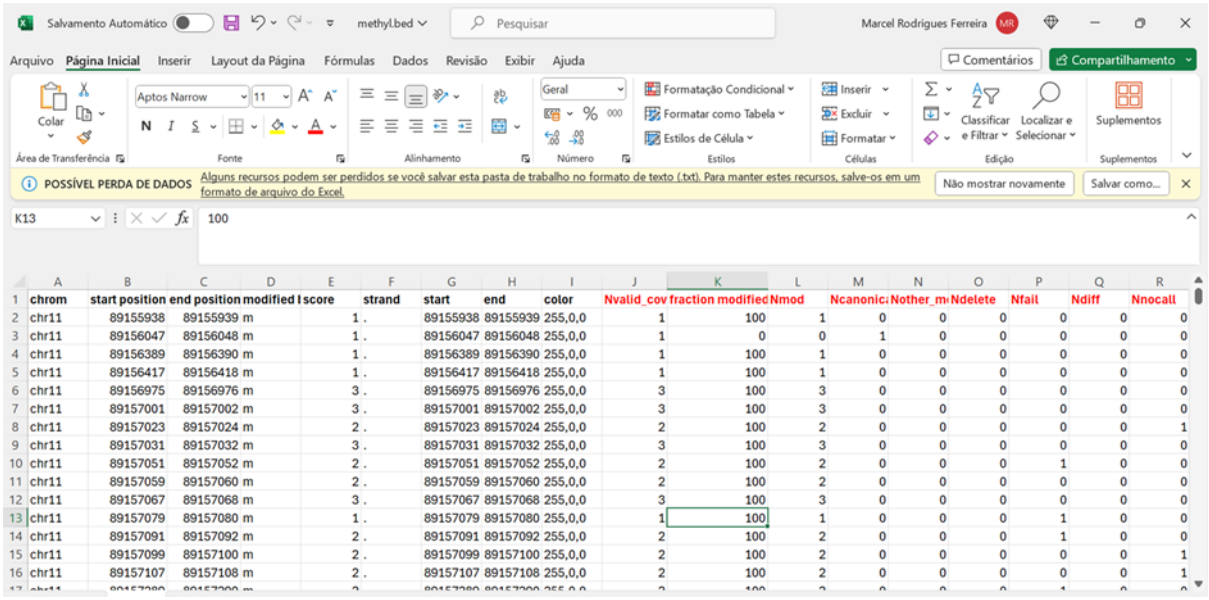
Para análise de metilação é necessário realizar a chamada de base utilizando configurações treinadas para detectar tais modificações. Como dito anteriormente, *dorado* é capaz de realizar a chamada de bases modificadas e salvar num arquivo BAM que possui duas *tags* (etiquetas) em seu cabeçalho para armazenar as informações dessas modificações (*Tags* MM e

ML). Este arquivo BAM deve então ser alinhado, via *dorado* ou *minimap2*, e por fim ordenado e indexado com *samtools*. Este arquivo então pode ser visualizado utilizando IGV como na figura abaixo. Este BAM pode ser analisado utilizando *modkit* (https://github.com/nanoporetech/modkit), que é uma software desenvolvido pela ONT para conversão em arquivos bedMethyl e calculos de estatisticas dessas modificações. O formato de arquivo bedMethyl contém informações essenciais sobre bases metiladas em uma sequência genômica. Cada linha da tabela representa um segmento específico da sequência, com colunas que detalham vários atributos. A coluna "***chrom***" indica o nome da sequência de referência do cabeçalho BAM (normalmente nome do cromossomo), enquanto "***start position***" e "***end position***" indicam as posições inicial e final exclusivas baseadas em 0, respectivamente, do segmento. A coluna "***modified base code and motif***" fornece um código de letra única para a base e o motivo modificados, com a "***pontuação***" representando a contagem de cobertura válida. A coluna "***strand***" se refere a fita do DNA entre valores positivos ("+") e negativos ("-"), ou "." quando as fitas são combinadas. Outras colunas, como '***color***' (sempre 255,0,0 para compatibilidade), 'N_{valid_cov}' (contagem de cobertura válida), 'fraction modified' (proporção de bases modificadas em relação à cobertura válida) e várias contagens, como 'N_{mod}', '$N_{canonical}$', 'N_{other_mod}', 'N_{delete}', 'N_{fail}', 'N_{diff}' e 'N_{nocall}', oferecem métricas detalhadas sobre os níveis de metilação e a qualidade do sequenciamento. *Modkit* descreve esses parametros como:

- N_{mod} - Número de chamadas que passaram pelos filtros e foram classificadas como um resíduo com uma modificação de base especificada.
- $N_{canonical}$ - Número de chamadas que passaram pelos filtros e foram classificadas como a base canônica em vez de modificada. A base exata deve ser inferida pelo código de modificação. Por exemplo, se o código de modificação for m (5mC), a base canônica será a citosina. Se o código de modificação for a, a base canônica será a adenina.
- $N_{other\ mod}$ - Número de chamadas que passaram pelos filtros e foram classificadas como modificadas, mas em que a modificação é diferente da base listada (e a base canônica correspondente é igual). Por exemplo, para uma determinada citosina, pode haver 3 leituras com h, 1 com uma identificação canônica e 2 com m identificações. Na linha bedMethyl para h, $\mathbf{N}_{\mathbf{other_mod}}$ seria 2. Na linha m, $\mathbf{N}_{\mathbf{other_mod}}$ seria 3.
- N_{valid_cov} - a cobertura válida. $N_{valid_cov} = N_{mod} + N_{other\ mod} + N_{canonical}$, também usado como a pontuação no bedMethyl
- N_{diff} - Número de leituras com uma base diferente da base canônica para essa modificação. Por exemplo, em uma linha para h a base canônica é citosina, se houver 2 leituras com substituições C->A, N_{diff} será 2.

- N_{delete} - Número de leituras com uma exclusão nessa posição de referência
- N_{fail} - Número de chamadas em que a probabilidade da chamada estava abaixo do limite. O limiar pode ser definido na linha de comando ou calculado a partir dos dados (geralmente falha no 10º percentil mais baixo de chamadas).
- N_{nocall} - Número de leituras alinhadas a essa posição de referência, com a base canônica correta, mas sem uma chamada de modificação de base. Isso pode ocorrer, por exemplo, se o modelo exigir um dinucleotídeo CpG e a leitura tiver uma substituição CG->CH, de modo que nenhuma chamada de modificação tenha sido produzida pelo analisador de bases.

Abaixo um exemplo de saída de um arquivo bedMethyl do *modkit* (**Figura 25**), importante destacar que como padrão as colunas N_{valid_cov} até N_{nocall} são colapsadas em uma única coluna sendo os valores separados por espaços.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	chrom	start position	end position	modified b	score	strand	start	end	color	Nvalid_cov	fraction modified	Nmod	Ncanonic	Nother_m	Ndelete	Nfail	Ndiff	Nnocall
2	chr11	89155938	89155939	m	1	.	89155938	89155939	255,0,0	1	100	1	0	0	0	0	0	0
3	chr11	89156047	89156048	m	1	.	89156047	89156048	255,0,0	1	0	0	1	0	0	0	0	0
4	chr11	89156389	89156390	m	1	.	89156389	89156390	255,0,0	1	100	1	0	0	0	0	0	0
5	chr11	89156417	89156418	m	1	.	89156417	89156418	255,0,0	1	100	1	0	0	0	0	0	0
6	chr11	89156975	89156976	m	3	.	89156975	89156976	255,0,0	3	100	3	0	0	0	0	0	0
7	chr11	89157001	89157002	m	3	.	89157001	89157002	255,0,0	3	100	3	0	0	0	0	0	0
8	chr11	89157023	89157024	m	2	.	89157023	89157024	255,0,0	2	100	2	0	0	0	0	0	1
9	chr11	89157031	89157032	m	3	.	89157031	89157032	255,0,0	3	100	3	0	0	0	0	0	0
10	chr11	89157051	89157052	m	2	.	89157051	89157052	255,0,0	2	100	2	0	0	0	1	0	0
11	chr11	89157059	89157060	m	2	.	89157059	89157060	255,0,0	2	100	2	0	0	0	0	0	0
12	chr11	89157067	89157068	m	3	.	89157067	89157068	255,0,0	3	100	3	0	0	0	0	0	0
13	chr11	89157079	89157080	m	1	.	89157079	89157080	255,0,0	1	100	1	0	0	0	1	0	0
14	chr11	89157091	89157092	m	2	.	89157091	89157092	255,0,0	2	100	2	0	0	0	1	0	0
15	chr11	89157099	89157100	m	2	.	89157099	89157100	255,0,0	2	100	2	0	0	0	0	0	1
16	chr11	89157107	89157108	m	2	.	89157107	89157108	255,0,0	2	100	2	0	0	0	0	0	1

Figura 25 - bedMethyl gerado por modkit utilizando o resultado de chamada de bases modificadas da amostra HG002 do GIAB. Nomes de colunas foram inseridos manualmente. As colunas marcadas em vermelho foram expandidas do seu formato original para melhor visualização. A coluna C (start position) é a posição da base modificada, devido ao formato de posição iniciando em 0.

Abaixo segue um fluxograma (**Figura 26**) para entendimento das principais etapas para análise de modificações de bases em dados de sequenciamento de Oxford Nanopore.

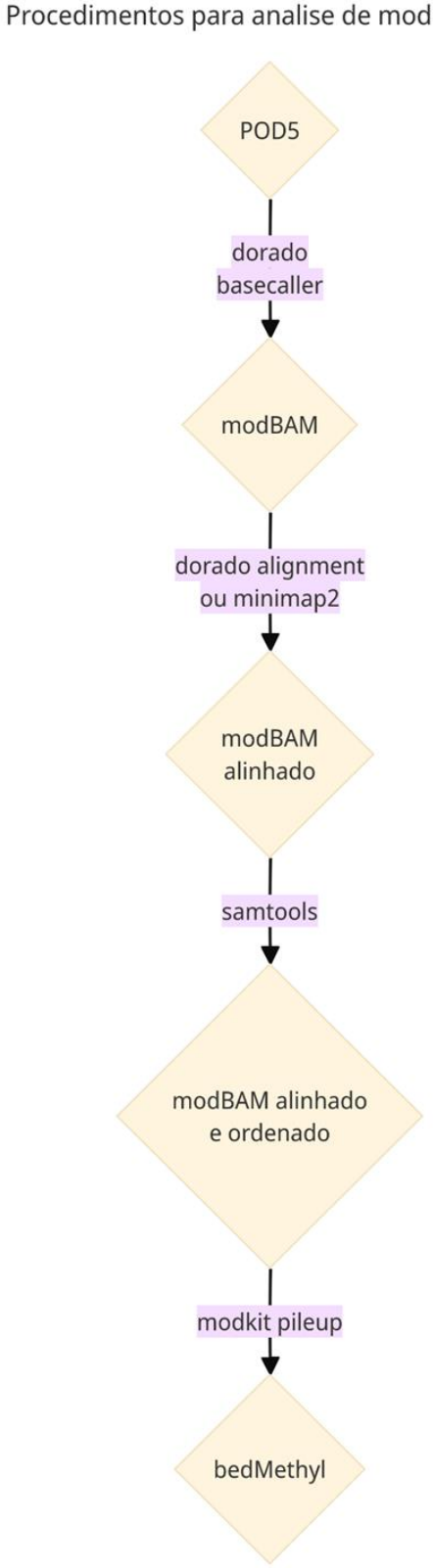


Figura 26 - Fluxograma de procedimentos para análise de modificações de bases em dados de sequenciamento Oxford Nanopore. É necessário utilizar uma configuração adequada para a detecção das modificações de interesse (Ex: 5mC ou 5hmC) já no processo de chamada de base.

Os dados foram gerados a partir da amostra HG002 do GIAB, sendo realizada a chamada de base com *dorado* incluindo a opção de detecção de metilação *5mCG_5hmCG* (recomendado pela ONT). O tempo total do procedimento foi de aproximadamente 5 dias. Ao final, os 1.3 Tb de arquivos POD5 geraram um BAM de aproximadamente 110 Gb ao ser alinhado e ordenado. Filtramos este BAM para as regiões dos genes de interesse e finalmente utilizamos o *modkit pileup* para gerar o bedMethyl da amostra. O tempo de execução do *modkit* foi de aproximadamente 1 minuto. Abaixo uma visualização da amostra no IGV para a região do gene MC1R (**Figura 27**). O arquivo bedMethyl também pode ser importado para o IGV (trilha superior) indicando regiões com detecção, neste caso, da metilação de citosinas (5mC). O IGV pode utilizar a informação contida no BAM modificado para colorir as reads como vistos em blocos vermelhos (posições com probabilidade superior a 50% para metilação) e azuis (posições com probabilidade inferior a 50% para metilação).

Figura 27 - Arquivo BAM com as modificações visualizado no IGV para a região do gene MC1R utilizando a amostra HG002 do GIAB. Foram selecionadas as opções de visualização color aligment by base modification e "*squished*".

Abaixo uma imagem destacando a diferença de informações que os arquivos BAM modificado e bedMethyl contém ao serem visualizados no IGV. bedMethyl fornece a informação do score da chamada de metilação para a base C enquanto o BAM contém informações de cada read para C e G (anotação da fita complementar) e apresenta sobre o total de reads e suas respectivas em C e G. *Modkit* consegue lidar com essa informação resumindo e indicando como modificações marcadas em C colapsando essa informação.

A figura abaixo (**Figura 28**) demonstra a diferentes informações que o IGV nos da sobre a metilação da amostra. A região analisada é chr16:89,919,709-89,919,759, gene MC1R, com os

arquivos do BAM modificado o do bedMethyl. Acima vemos o painel de visualização quando clicamos no bedMethyl indicando o nome da modificação (m de metilação), a localização no genoma e um valor de score. Este score representa na prática quantas vezes houve leitura nesta posição e não um indício do nível de metilação. Na sequência temos os painéis de cobertura da metilação. A esquerda da fita + e a direita da fita -. Estes exibem informações sobre o total de cobertura número de leituras das bases e a probabilidade de modificação para a determinada base, no caso do painel da direita podemos ver inclusive uma leitura de modificação 5hmC com 60% de probabilidade enquanto temos 22 bases com modificação 5mC com 98% de probabilidade para esta modificação. Por fim temos os painéis para as leituras simples que contém as probabilidades de modificação para aquela leitura. A direita temos uma leitura em roxo que contém 60% de probabilidade de ser 5hmC e 38% de ser 5mC.

Figura 28 - Arquivo BAM com as modificações visualizado no IGV para a região do gene MC1R (chr16:89,919,709-89,919,759) utilizando a amostra HG002 do GIAB. Paineis de visualização do bedMethyl, cobertura do BAM (ambas as fitas) e duas bases diferentes (uma de cada fita).

4.3 Análises bioinformáticas em dados públicos do Projeto 1000 Genomes

Após a definição e validação dos procedimentos computacionais com a amostra de referência HG002, aplicamos o pipeline de bioinformática às 57 amostras públicas do Projeto 1000 Genomes, geradas com a tecnologia Oxford Nanopore (ONT). Essas amostras, oriundas de diferentes populações e sequenciadas com as químicas R10 e protocolos de *long* e *ultra long reads*, foram analisadas com foco na identificação de variantes genômicas e no perfil de metilação. Esta etapa teve como objetivo avaliar a robustez e escalabilidade dos métodos estabelecidos, bem como investigar a variabilidade genômica e epigenética entre indivíduos de distintas origens populacionais.

4.3.1 Análise da cobertura dos genes

Para avaliar a cobertura de cada gene em cada amostra, utilizamos a ferramenta *samtools coverage*, que permitiu calcular a profundidade média de leitura ao longo das regiões gênicas previamente definidas no arquivo bed. A **Figura 29** apresenta um mapa de calor ilustrando a distribuição da cobertura média para cada gene nas diferentes amostras analisadas. A escala de cores reflete a profundidade média de leitura, variando de tons mais escuros, que indicam menor cobertura, a tons mais claros, representando regiões com maior profundidade de sequenciamento. Observa-se uma variação significativa na cobertura entre os genes e entre as amostras, sugerindo possíveis diferenças na eficiência do sequenciamento, no mapeamento das leituras ou nas características genômicas específicas das regiões analisadas. Alguns genes, como ASIP, KITLG e MC1R, apresentam cobertura relativamente alta em várias amostras, enquanto outros exibem padrões mais heterogêneos, possivelmente refletindo flutuações na qualidade do sequenciamento ou na representação dessas regiões nos dados analisados. Os genes com menor cobertura pode ser identificados pelas regiões mais escuras no *heatmap*, indicando profundidade de leitura reduzida. Observando a figura, os genes TPCN2, TYR e TYRP1 apresentam algumas das coberturas mais baixas na maioria das amostras, conforme evidenciado pelas tonalidades mais escuras. Isso sugere que essas regiões podem ter sido menos eficientemente sequenciadas ou mapeadas, possivelmente devido a fatores como alta complexidade genômica, baixa eficiência de captura ou dificuldades inerentes ao alinhamento nessas regiões específicas.

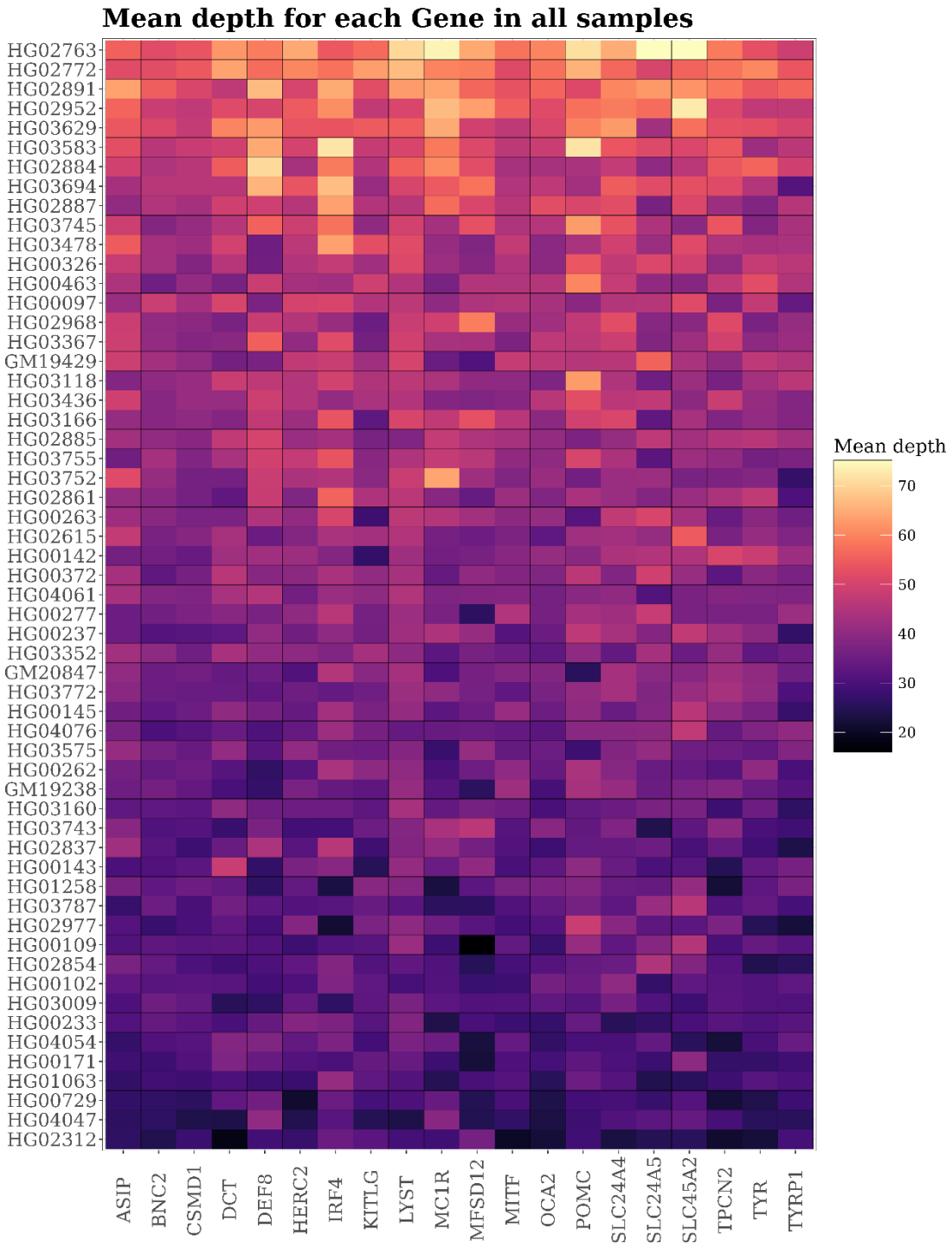


Figura 29 - Cobertura média por gene em cada amostra. Amostras foram ordenadas da maior para menor cobertura.

A **Figura 30** contém o gráfico de densidade exibindo a distribuição dos valores de cobertura média para cada amostra analisada. A densidade indica como os valores de cobertura estão distribuídos dentro de cada amostra, permitindo avaliar a homogeneidade da profundidade de sequenciamento entre os genes. As cores variam de tons escuros, representando coberturas mais baixas, até tons claros, indicando coberturas mais elevadas. A maioria das amostras apresenta

uma cobertura média dos genes concentrada entre 20 e 50, com algumas atingindo valores superiores a 70. Amostras como HG02763, HG02772 e HG02891 demonstram uma cobertura mais elevada, com distribuições deslocadas para valores médios mais altos e com maior intensidade em tons claros. Por outro lado, amostras como HG04047, HG02312 e HG01063 exibem uma cobertura média menor, com distribuições mais concentradas em valores abaixo de 30, indicando uma menor profundidade de sequenciamento destes genes. A baixa cobertura observada em algumas amostras pode comprometer análises subsequentes, como chamadas de variantes e inferências funcionais, uma vez que profundidades insuficientes aumentam a chance de erros e reduzem a confiabilidade dos dados genômicos obtidos.

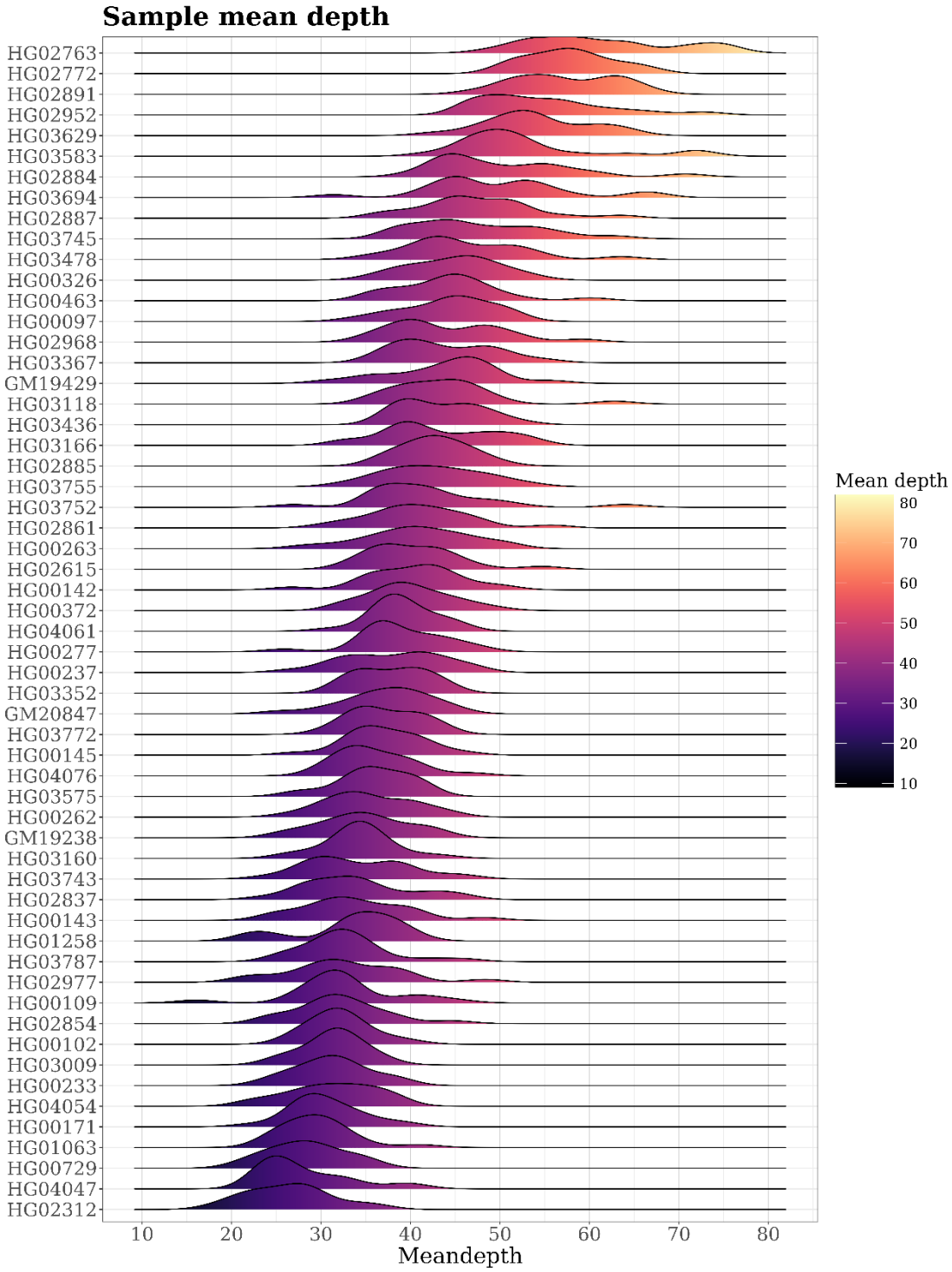


Figura 30 - Distribuição de valores de cobertura média por amostra. Amostras foram ordenadas da maior para menor cobertura.

4.3.2 Haplótipos longos dos genes de pigmentação

A fim de aprofundar nossa compreensão sobre a diversidade alélica dos genes de pigmentação, realizamos a montagem de haplótipos longos para cada gene de interesse em diferentes indivíduos. Utilizamos o software PECAT, que integra um pipeline completo para montagem genômica de novo com resolução de haplótipos a partir de dados de sequenciamento

de leitura longa. O PECAT (https://github.com/lemene/PECAT) emprega ferramentas como Samtools (https://github.com/samtools/samtools), Clair3 (https://github.com/HKU-BAL/Clair3) e Medaka (https://github.com/nanoporetech/medaka) para chamada e polimento de variantes, permitindo o agrupamento preciso das leituras por haplótipo e a posterior montagem individual de cada um. Essa abordagem foi pensada com o objetivo de recuperar ambos os haplótipos de cada gene por indivíduo, fornecendo uma base robusta para análises comparativas estruturais e funcionais que poderiam ser mascaradas em montagens colapsadas ou consenso.

A **Figura 31** apresenta a distribuição das amostras de acordo com o número de haplótipos detectados para cada um dos 20 genes de pigmentação avaliados. Observa-se uma variabilidade considerável na eficiência da montagem haplotípica entre os diferentes genes. Enquanto genes como ASIP, SLC45A2, TYRP1 e TYR apresentaram dois haplótipos em praticamente todas as 57 amostras, outros, como MC1R, DCT, HERC2 e MITF, apresentaram um número significativo de casos com apenas um ou nenhum haplótipo montado. Essa variação pode estar relacionada a características genômicas específicas de cada gene, como a presença de regiões altamente repetitivas, polimorfismos estruturais ou até mesmo baixa cobertura de leitura em determinadas amostras. Esses resultados indicam que, embora a montagem haplotípica com PECAT tenha sido bem-sucedida para a maioria dos genes e amostras, há uma limitação na recuperação de ambos os haplótipos em alguns contextos, o que deve ser considerado nas análises subsequentes.

A **Figura 32** mostra o número de contigs encontrados para cada haplótipo dos genes analisados nas 57 amostras. Observa-se que, embora muitos haplótipos tenham sido reconstruídos em um único bloco contíguo (1 contig), há diversos casos em que a montagem resultou em múltiplos contigs para um mesmo haplótipo. Essa fragmentação sugere que, para determinadas combinações de amostra e gene, a montagem não conseguiu abranger todo o alelo de forma contínua, o que pode ser consequência de regiões com baixa cobertura, alta complexidade genômica ou presença de variantes estruturais. Esse padrão é particularmente evidente em genes como HERC2, DCT e MITF, nos quais alguns haplótipos foram representados por até 3 ou mais contigs. Esses resultados indicam limitações na completude da montagem haplotípica e reforçam a necessidade de considerar a qualidade estrutural dos haplótipos antes de análises funcionais ou comparativas.

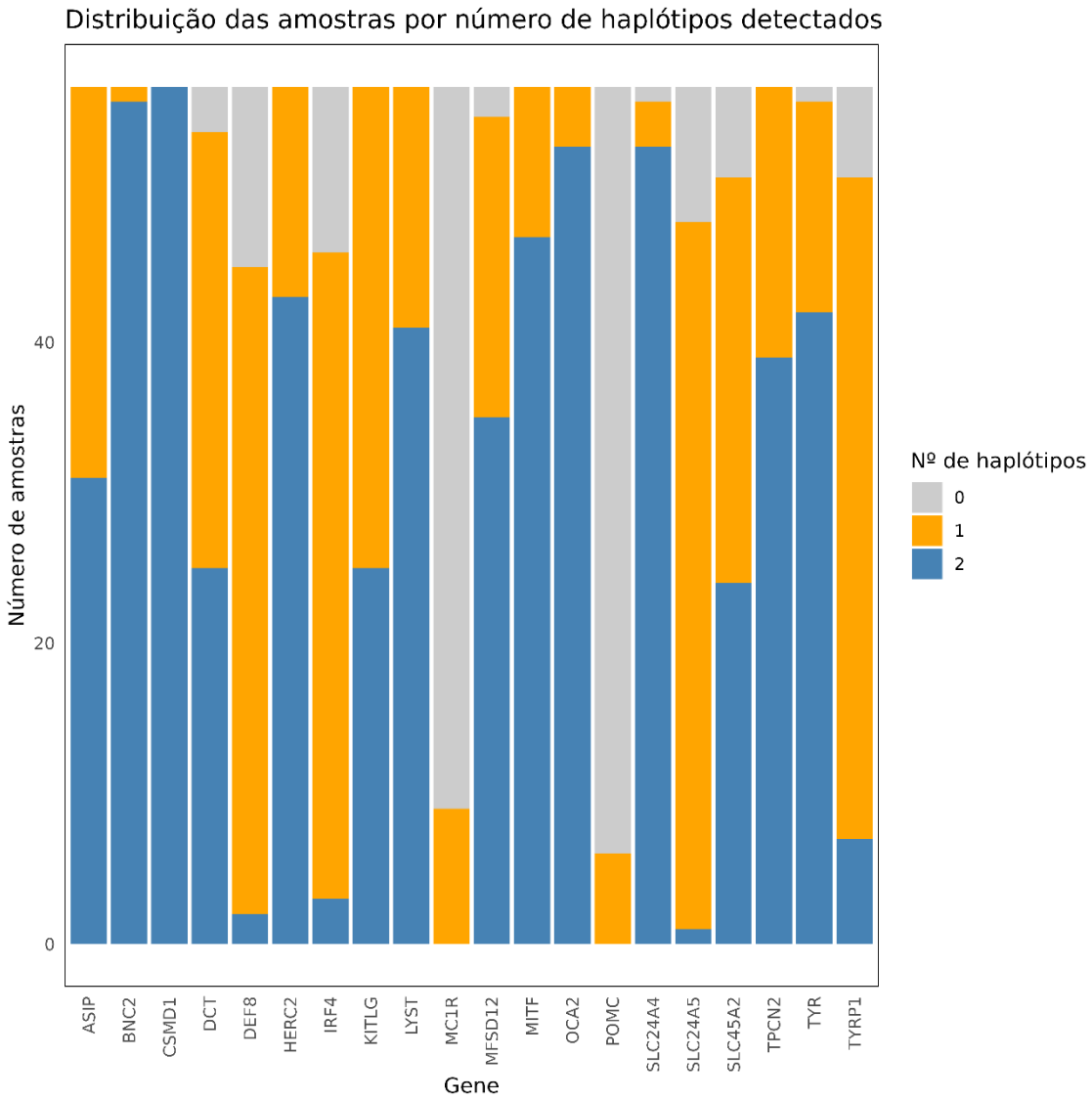


Figura 31 - Número de haplótipos encontrados por gene/amostra.

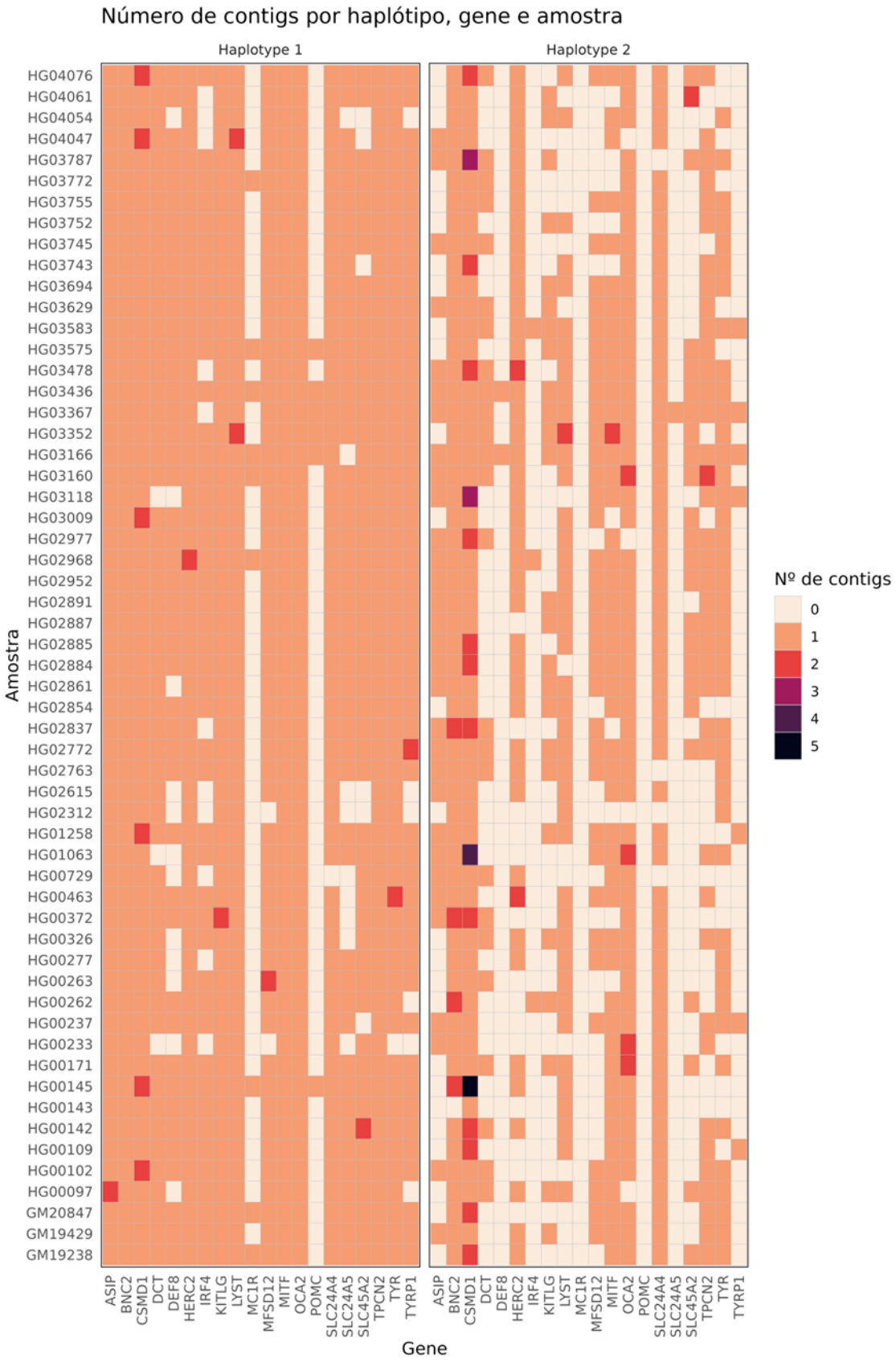


Figura 32 - Número de contigs encontrados para cada haplótipo dos genes nas 57 amostras analisadas.

A **Figura 33** apresenta a proporção da região do gene coberta por cada haplótipo, considerando todos os contigs encontrados nas 57 amostras analisadas. Para isso, os haplótipos reconstruídos foram alinhados ao genoma de referência e, em seguida, a profundidade de cobertura foi calculada com a ferramenta *samtools depth*. A porcentagem final representa o quanto da extensão esperada do gene está coberta por pelo menos uma base dos contigs do haplótipo. Observa-se que a maioria dos haplótipos apresenta alta cobertura dos genes, com valores próximos a 100% (cor vermelha escura). No entanto, há diversos casos com cobertura incompleta, refletidos por tons mais claros (laranja, amarelo, verde ou azul), indicando regiões parcialmente ausentes nos haplótipos reconstruídos. Esses casos ocorrem em diferentes genes e amostras, com maior frequência no Haplótipo 2, sugerindo possível assimetria na qualidade ou completude da montagem entre os dois alelos. Esse resultado complementa a **Figura 32**, mostrando que a fragmentação dos haplótipos em múltiplos contigs não comprometeu, na maioria dos casos, a cobertura completa da região gênica. Ainda assim, a ausência de cobertura total em alguns haplótipos reforça a necessidade de considerar essas limitações antes de inferir variações funcionais ou estruturais a partir desses dados.

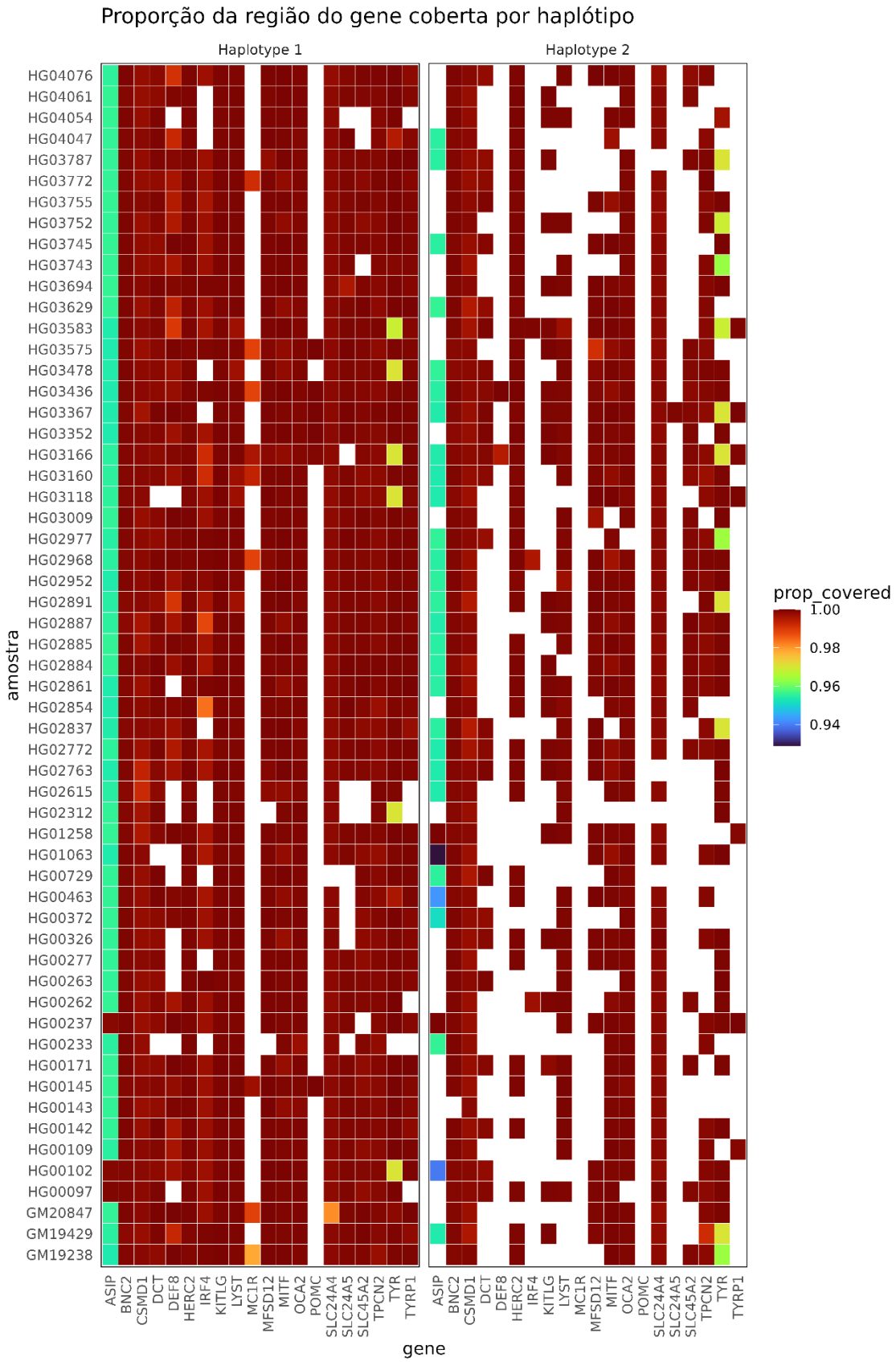


Figura 33 - Porcentagem coberta do gene pelos haplótipos (considerando todos os contigs) encontrados.

A seguir, apresento dois casos particularmente interessantes observados na análise de cobertura haplotípica. A **Figura 34** detalha o primeiro deles: a cobertura por posição para os dois haplótipos do gene ASIP nas diferentes amostras analisadas. Cada linha representa uma amostra individual, com os haplótipos 1 e 2 exibidos separadamente, em azul e amarelo, respectivamente. As faixas coloridas indicam as regiões do gene que foram cobertas por contigs, enquanto a linha na parte inferior do gráfico mostra os éxons anotados no gene de referência, permitindo uma comparação direta entre cobertura e estrutura gênica. Esse painel complementa as observações feitas nas Figuras 32 e 33. A Figura 32 destacou o número de contigs por haplótipo, evidenciando uma montagem frequentemente fragmentada. Já a Figura 33 apresentou a proporção de cobertura ao longo do gene, sugerindo uma alta cobertura em muitos haplótipos, embora com lacunas. A Figura 34 aprofunda essa análise ao revelar que, mesmo quando a cobertura total parece satisfatória, há regiões específicas do gene ASIP que permanecem descobertas em diversos haplótipos — especialmente no haplótipo 2, onde interrupções recorrentes são visíveis. Notavelmente, algumas dessas falhas coincidem com regiões exônicas, o que pode impactar análises funcionais e inferência de variantes. Esse padrão assimétrico entre os haplótipos destaca a importância de análises por posição para genes de interesse, e reforça a necessidade de avaliar os alelos separadamente. As falhas na cobertura podem refletir limitações técnicas na montagem, variabilidade estrutural entre haplótipos ou até diferenças biológicas relevantes, exigindo uma investigação mais cuidadosa.

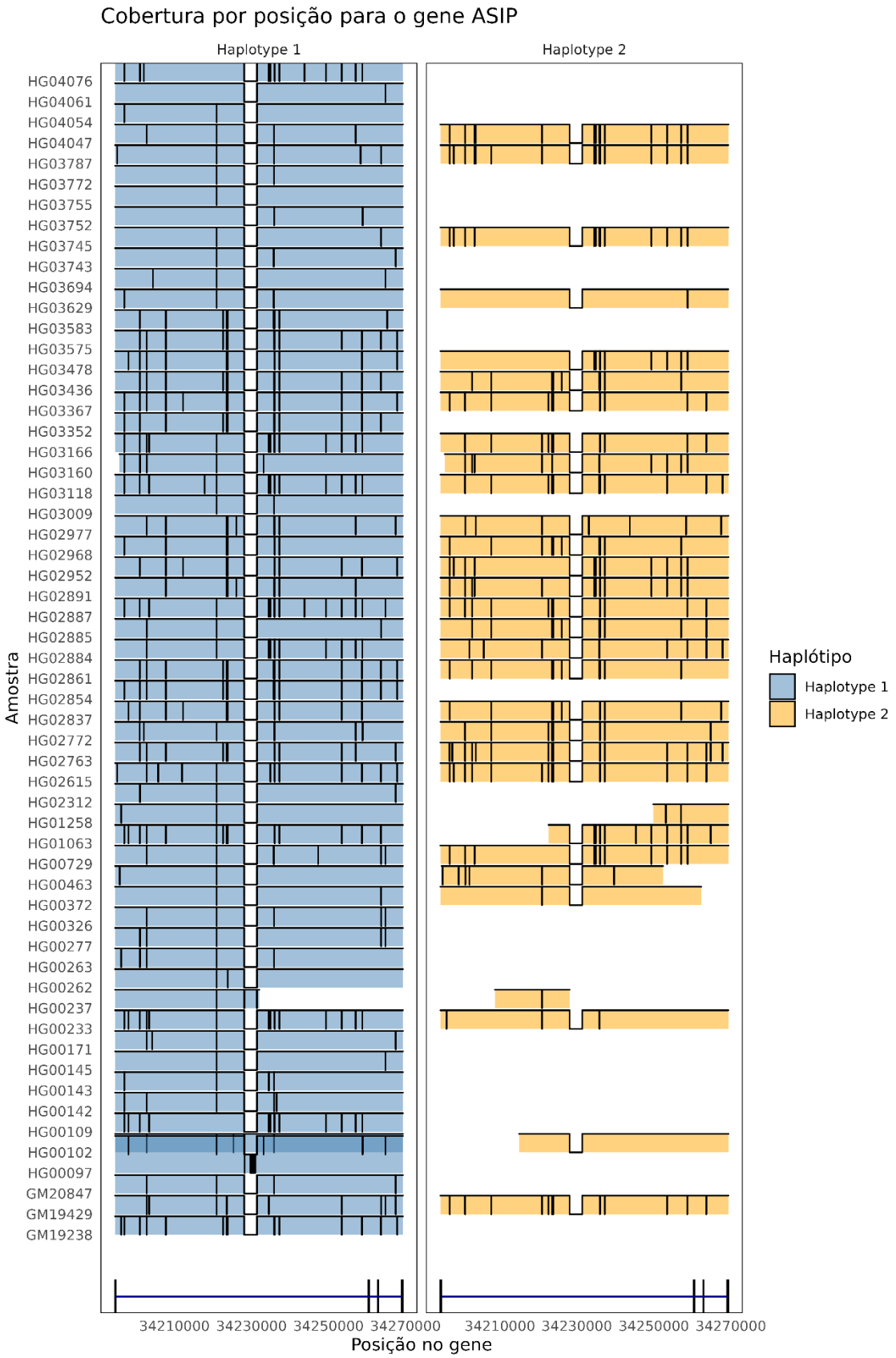


Figura 34 - Cobertura dos haplótipos de ASIP para as diferentes amostras analisadas. Abaixo contem um inserto representando os exóns do gene.

O segundo caso interessante observado envolve o gene DEF8, cuja cobertura haplotípica está ilustrada na **Figura 35**. Um padrão marcante neste gene é que, na grande maioria das amostras, apenas um dos haplótipos foi montado — o que já sugere uma dificuldade na resolução haplotípica completa para essa região. Além disso, chama a atenção a grande variação no comprimento dos haplótipos montados: alguns contigs são curtos e cobrem apenas partes do gene, enquanto outros se estendem por regiões muito maiores. Essa observação nos levou a investigar os genes vizinhos, TUBB3 e MC1R, também destacados na parte inferior da figura. Foi então que percebemos que diversos contigs atribuídos ao haplótipo de DEF8 estavam, na verdade, cobrindo simultaneamente os três genes — um indicativo de que, em alguns casos, a montagem uniu regiões genômicas próximas em um único haplótipo estendido. Esse achado é particularmente relevante para o estudo de MC1R, um gene de interesse funcional, para o qual não conseguimos obter haplótipos completos na maioria das amostras. A sobreposição entre os haplótipos de DEF8 e as regiões de MC1R abre uma possibilidade interessante: combinar as reads de DEF8 com as de MC1R em uma estratégia conjunta de montagem. Isso pode aumentar a probabilidade de reconstrução haplotípica completa para MC1R, melhorando a cobertura e permitindo análises mais robustas sobre variação alélica e estrutura gênica.

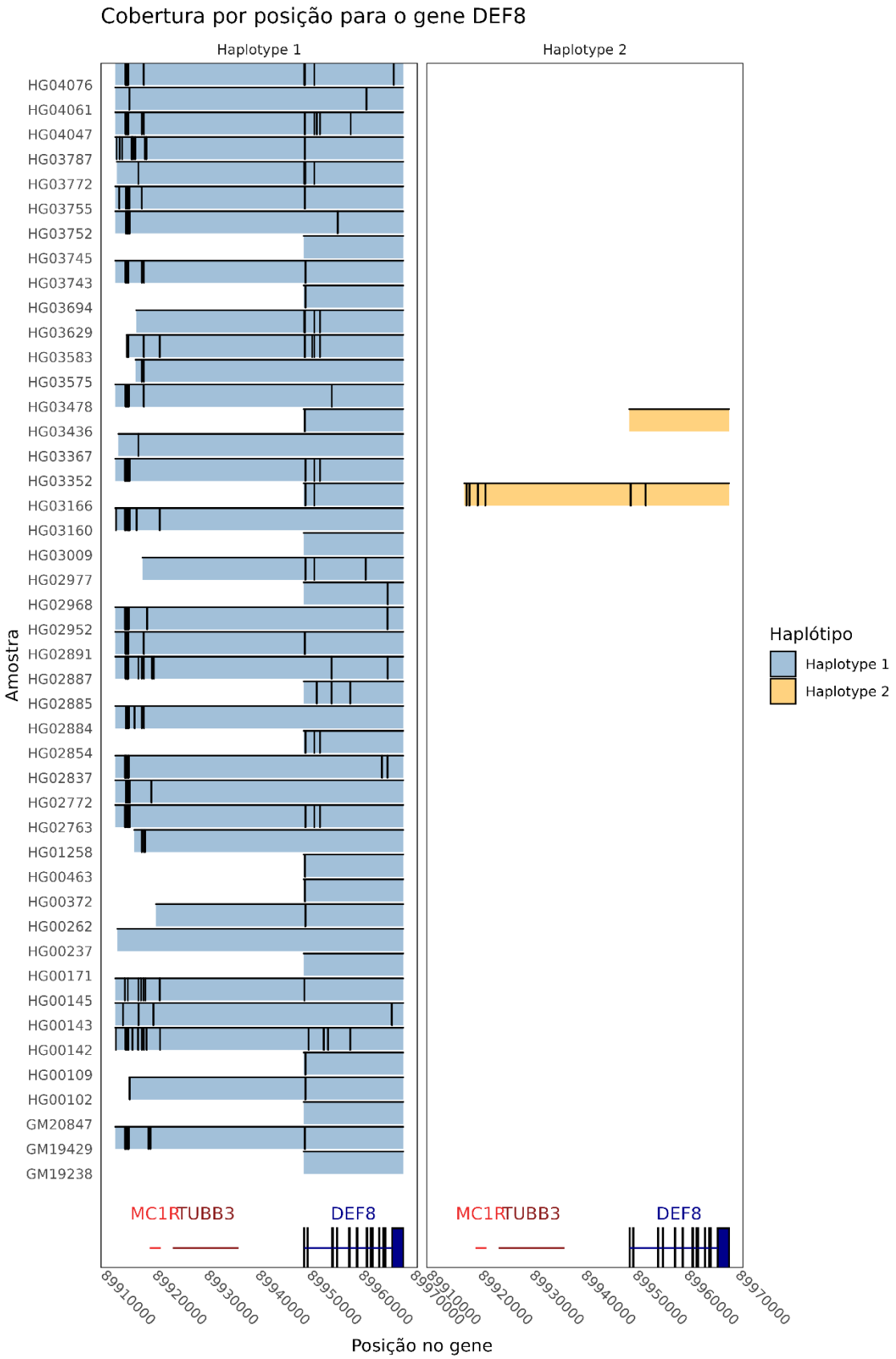


Figura 35 - Cobertura dos haplótipos de DEF8 para as diferentes amostras analisadas .

Os resultados obtidos através da análise de haplótipos estendidos utilizando a ferramenta PECAT demonstraram-se interessantes, fornecendo insights valiosos sobre a estrutura haplotípica e a diversidade genômica investigada. No entanto, ficou evidente que a otimização dos parâmetros de análise é crucial para melhorar a precisão e a resolução dos dados, uma vez que variações nas configurações podem impactar significativamente a detecção de blocos de haplótipos e a concordância com dados experimentais. Adicionalmente, a utilização de outras ferramentas computacionais, como Hifiasm (Cheng et al., 2021) ou Shasta-Hapdup (Shafin et al., 2020), pode representar uma estratégia promissora para refinamento dos resultados. O Hifiasm, por exemplo, é reconhecido por sua eficiência na montagem de novo de genomas diploides, enquanto o Shasta-Hapdup oferece uma abordagem robusta para a resolução de haplótipos a partir de sequenciamento de leitura longa. Esses métodos devem ser empregados em trabalhos futuros.

4.3.3 Metilação de genes de pigmentação

Para investigar os padrões epigenéticos associados aos genes de pigmentação, foi utilizado o *software* Modkit, especificamente o comando *modkit pileup*, para processar os dados de metilação de 20 genes candidatos em 57 amostras. Essa abordagem resultou na geração de 1140 arquivos methylbed, que contêm as chamadas de metilação por posição ao longo dos genes. Esses arquivos foram utilizados na análise de regiões diferencialmente metiladas (DMRs) apresentada nas seções seguintes, com o objetivo de identificar regiões potencialmente associadas a diferenças fenotípicas de pigmentação. Para ilustrar o tipo de informação gerada a partir dos arquivos methylbed, a Figura 4.3 apresenta o perfil de metilação ao longo do gene SLC24A4 na amostra GM19238. Cada ponto representa uma posição metilada, com o eixo y indicando o beta-value, ou seja, a proporção de leituras modificadas em relação ao total de leituras naquela posição. O tamanho dos pontos reflete o número de eventos de metilação detectados (Nmod), enquanto a linha vermelha representa um modelo suavizado ajustado por GAM (*Generalized Additive Model*), permitindo visualizar a tendência da metilação ao longo do gene. Esse tipo de visualização é útil para identificar padrões locais de hipermetilação ou hipometilação, que podem estar associados à regulação gênica e à expressão diferencial dos genes de pigmentação entre os indivíduos analisados.

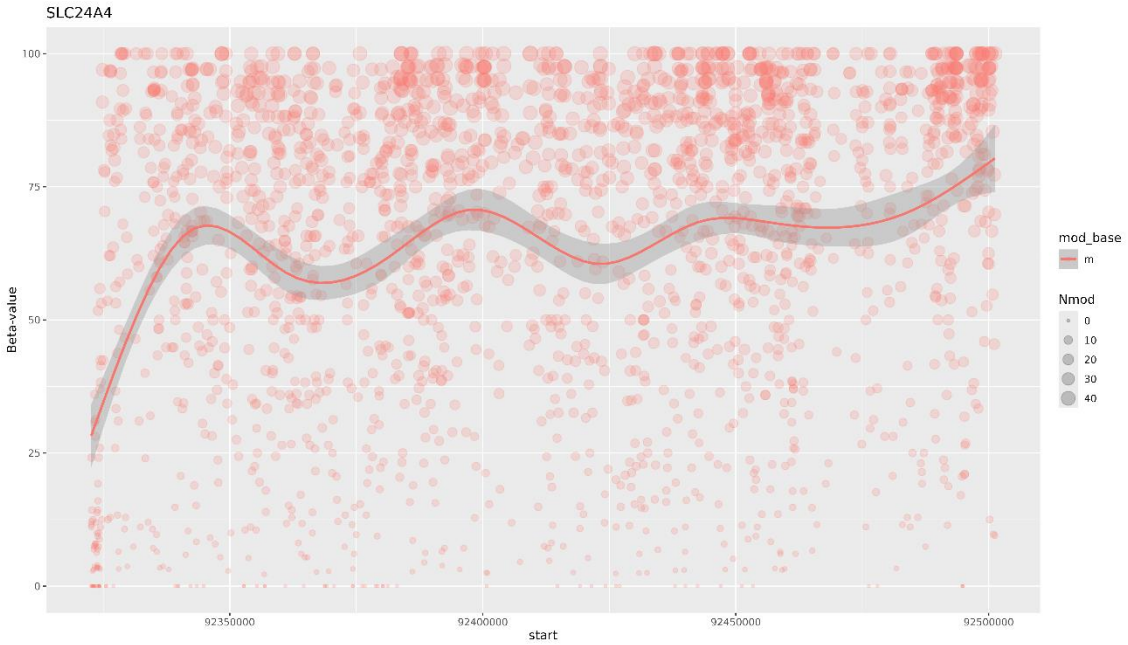


Figura 36 - Perfil de metilação do gene SLC24A4 na amostra GM19238.

4.3.4 Metilação diferencial entre haplótipos

Para a análise de regiões diferencialmente metiladas (DMRs), optamos por uma abordagem intraindividual, explorando as diferenças entre os haplótipos identificados em cada amostra. Como não havia grupos experimentais ou populacionais definidos, utilizamos a ferramenta LongPhase para a faseamento dos dados, permitindo a separação das leituras de cada haplótipo. Em seguida, aplicamos o módulo dmr do Modkit para identificar regiões do genoma que apresentaram diferenças significativas nos níveis de metilação entre os haplótipos. Essa estratégia possibilitou a investigação de assinaturas epigenéticas haplótipo-específicas, ampliando a resolução da análise e revelando possíveis efeitos alélicos na metilação do DNA.

A **Figura 37** contém um *heatmap* representando o número de citosinas diferencialmente modificadas (escala logarítmica base 10) entre haplótipos, em genes selecionados associados à pigmentação. Cada linha representa um indivíduo, e cada coluna corresponde a um gene. A intensidade da cor indica o número de modificações diferenciais detectadas por gene entre os dois haplótipos de cada indivíduo, conforme identificado pelo Modkit. Observa-se variação interindividual na carga de modificações diferenciais, com genes como CSMD1, BNC2 e HERC2 apresentando elevada densidade de modificações em diversos indivíduos.

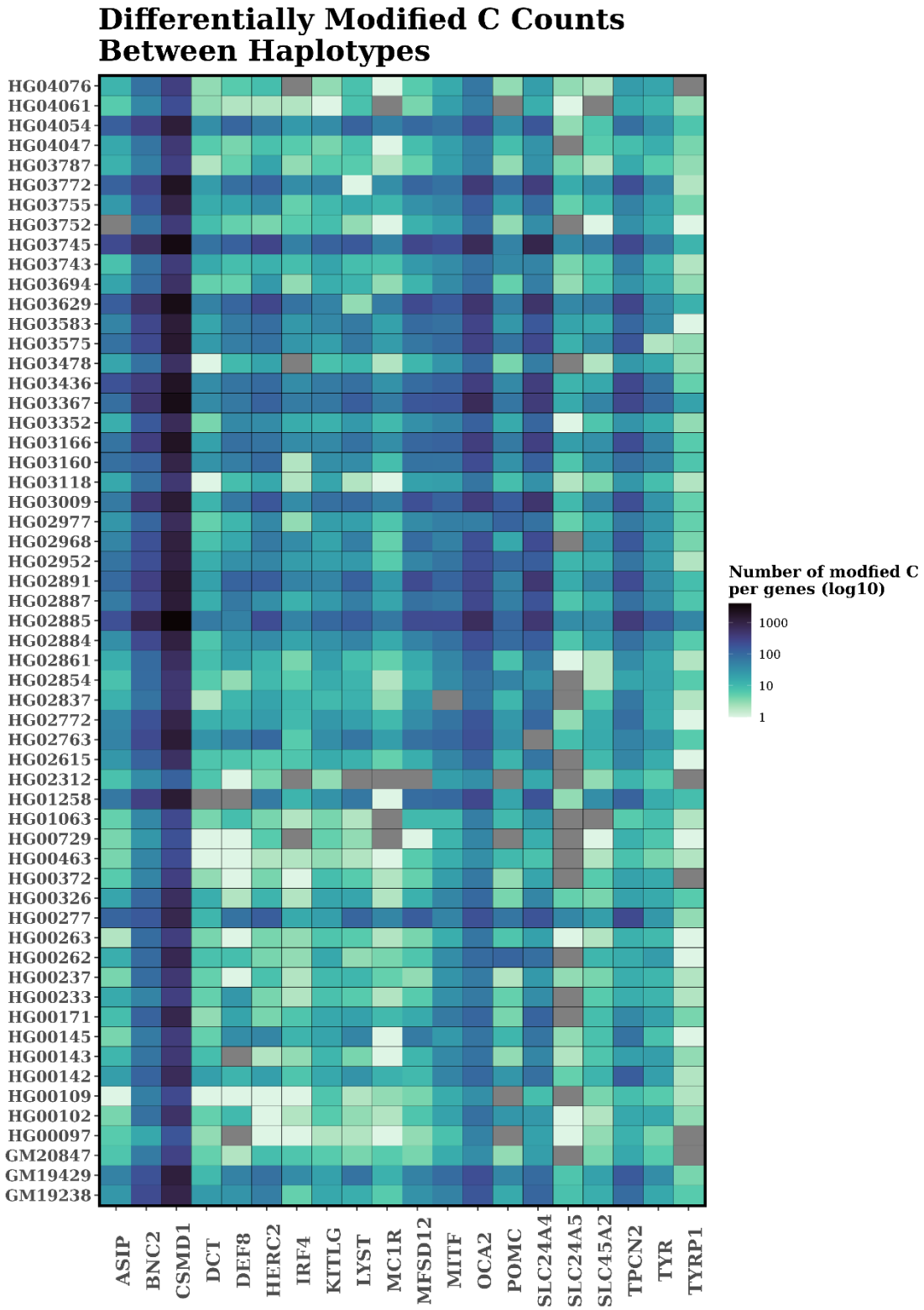


Figura 37 - Regiões diferencialmente metiladas (DMR) comparando-se os dois haplótipos de cada indivíduo/gene. Contagens apresentadas com escala log10.

Para interpretar os achados obtidos na análise genômica, investigamos mais detalhadamente o gene CSMD1, um dos genes com maior número de citosinas diferencialmente

modificadas entre haplótipos, conforme ilustrado na **Figura 37**. Utilizando o Integrative Genomics Viewer (IGV), foi possível visualizar, na região genômica chr8:3,744,135-3,744,216, a presença de três diferentes citosinas com metilação diferencial entre os dois haplótipos do indivíduo HG02885. Conforme mostrado na **Figura 38**, essas modificações ocorrem em uma região intrônica do gene CSMD1, que pode desempenhar papel regulatório. A observação dessas modificações alélicas reforça a hipótese de que a metilação diferencial entre haplótipos pode influenciar a regulação gênica de forma haplótipo-específica, sendo relevante para fenótipos associados à expressão diferencial de CSMD1.

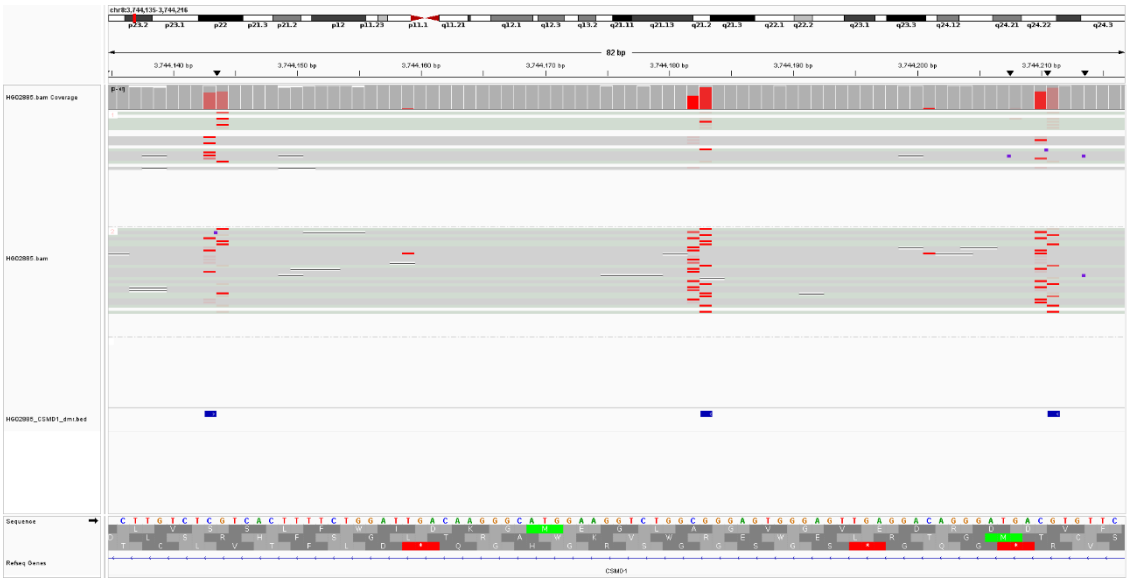


Figura 38 - Screenshot do IGV na região chr8:3,744,135-3,744,216 do gene CSMD1. A região contém exemplo de 3 diferentes citosinas com metilação diferencial entre os haplótipos do indivíduo HG02885.

Para investigar se o padrão de metilação observado em CSMD1 no indivíduo HG02885 era consistente em outros indivíduos, analisamos os dados de metilação da amostra HG00262, selecionada por apresentar ampla cobertura na região deste gene. A **Figura 39** mostra a distribuição dos valores beta ao longo do gene CSMD1, com cada ponto representando uma citosina metilada na base C. Observa-se que a metilação é predominantemente alta ao longo de todo o gene, com valores beta concentrados próximos de 100%. Essa metilação elevada é consistente com um estado epigenético repressivo típico de genes com expressão regulada espacial ou temporalmente. A presença generalizada dessa marca epigenética sugere que CSMD1 pode estar sob controle transcricional rigoroso, possivelmente modulado por mecanismos haplótipo-específicos como os visualizados na amostra HG02885.

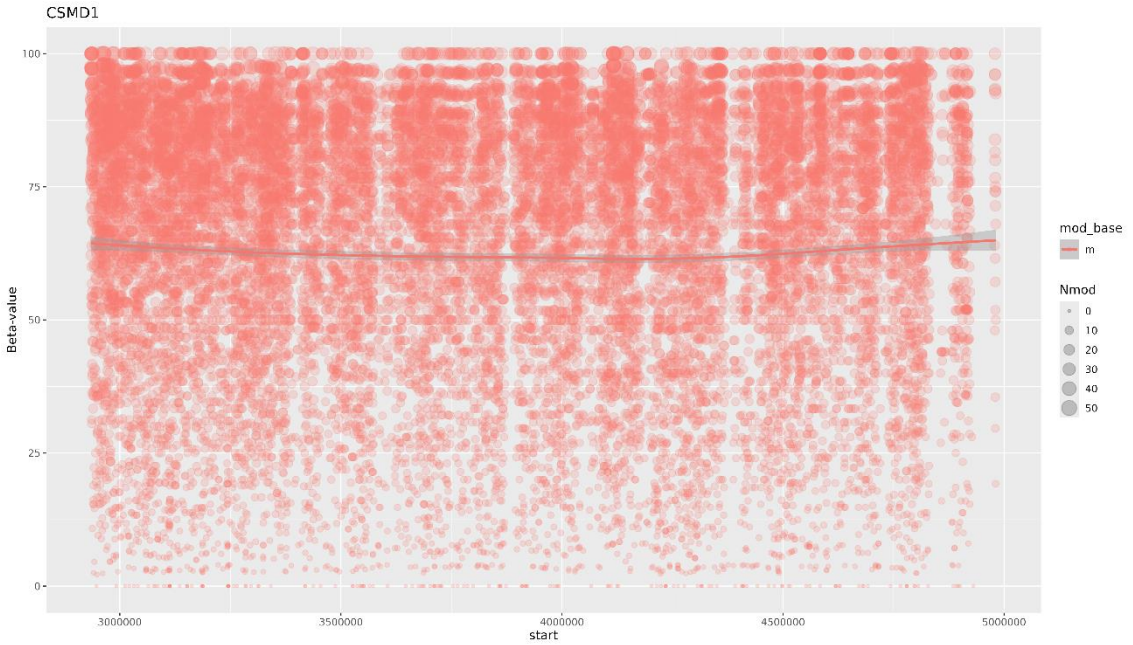


Figura 39 - Metilação ao longo do gene CSMD1. Amostra HG00262.

As análises de DMRs em nossas próprias amostras — coletadas e sequenciadas especificamente para este estudo — permitirão investigar tanto diferenças epigenéticas entre tecidos de um mesmo doador, quanto variações entre diferentes doadores ao se comparar tecidos equivalentes. Essa abordagem fornecerá uma visão mais abrangente do papel da metilação do DNA na regulação gênica associada à pigmentação. Ao integrar essas comparações, poderemos identificar assinaturas epigenéticas específicas de tecidos e potenciais variações interindividuais, contribuindo para o entendimento dos mecanismos regulatórios que influenciam fenótipos pigmentares humanos.

4.3.5 Entropia de metilação

A análise da entropia da metilação do DNA fornece uma medida da variabilidade epigenética em diferentes regiões do genoma, refletindo padrões dinâmicos de regulação gênica, estabilidade epigenética e heterogeneidade celular (Lee et al., 2023; Xie et al., 2011). Valores elevados de entropia podem indicar a coexistência de diferentes estados de metilação dentro da população celular, possivelmente associada a processos de diferenciação, plasticidade epigenética ou desorganização em condições patológicas, como o câncer. Por outro lado, baixa entropia sugere padrões de metilação mais estáveis, característicos de regiões funcionalmente reguladas. O cálculo da entropia é realizado utilizando-se a equação de entropia com fórmula abaixo, onde N é o conjunto de todos os padrões de metilação e Pr(ni) é a probabilidade empírica desse padrão.

$$ME = \frac{-1}{N}\sum_{N} Pr(n_i) * log_2\big(Pr(n_i)\big)$$

A **Figura 40** (disponibilizada pela ONT para explicar o conceito de entropia de metilação) ilustra o conceito de entropia de metilação (ME, Methylation Entropy), uma métrica quantitativa que mede a heterogeneidade dos padrões de metilação em uma população celular. Valores de ME variam entre 0 e 1, onde: ME = 0 indica metilação completamente homogênea (todos as moléculas apresentam o mesmo padrão de metilação), ME próximo a 1[2]: Reflete alta diversidade de padrões de metilação (heterogeneidade máxima). Podemos ver exemplos de cálculos de ME para diferentes cenários, como: ME = 0.25 (Baixa heterogeneidade), ME = 0.5 (Heterogeneidade moderada), e ME = 0.75 (Alta heterogeneidade).

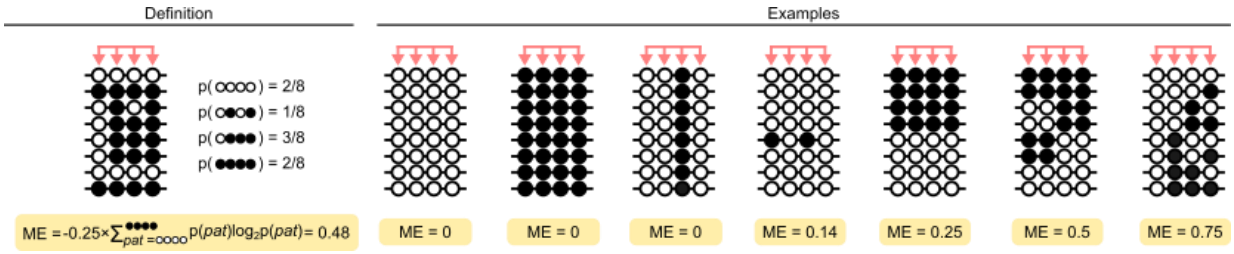


Figura 40 – Esquema sobre entropia de metilação. Figura extraída de https://nanoporetech.github.io/modkit/intro_entropy.html.

A análise de entropia de metilação revelou uma considerável variabilidade nos padrões epigenéticos entre os indivíduos e genes analisados. Observa-se que alguns genes apresentam valores consistentemente mais altos de entropia, indicando uma regulação epigenética mais dinâmica, enquanto outros exibem padrões de metilação mais conservados. Genes como POMC, SLC24A4 e MC1R apresentam regiões com alta entropia em diversos indivíduos, sugerindo que a metilação nesses loci é altamente variável entre os alelos ou haplótipos. Essa característica pode refletir um controle epigenético mais flexível, potencialmente associado a variações na expressão gênica.

Por outro lado, genes como TYR, ASIP e KITLG demonstram predominantemente valores baixos de entropia, indicando que seus padrões de metilação são mais estáveis e menos propensos a variações interindividuais. Isso sugere que a regulação epigenética dessas regiões pode ser mais conservada, o que pode estar relacionado à importância funcional desses genes e à necessidade de manter sua expressão dentro de limites mais rígidos.

Além disso, observa-se que certos indivíduos apresentam padrões de entropia distintos, o que pode refletir diferenças genéticas ou a influência de fatores ambientais na regulação epigenética. Enquanto alguns indivíduos possuem perfis de metilação mais homogêneos, outros exibem uma variabilidade considerável entre diferentes genes. No conjunto, os resultados indicam que a entropia da metilação não é uniformemente distribuída entre os genes nem entre os

[2] Entropia de metilação pode apresentar valores superiores a 1.

indivíduos, sugerindo que determinados loci podem estar sujeitos a um controle epigenético mais dinâmico. Essa heterogeneidade pode ter implicações funcionais na regulação da expressão gênica e no potencial impacto fenotípico da metilação do DNA. A **Figura 41** contém os valores da mediana da entropia de metilação para todos os genes e amostras.

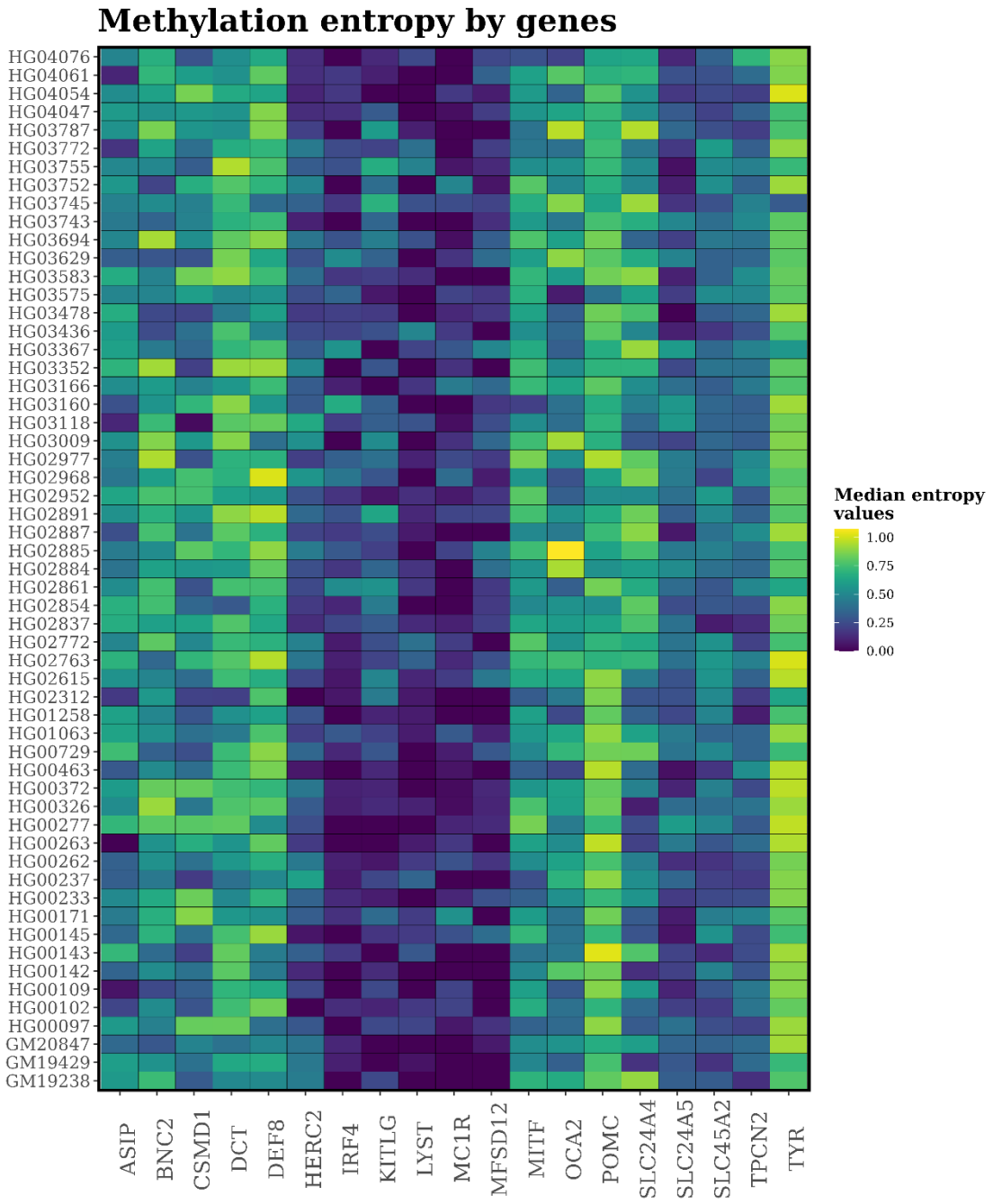


Figura 41 - Heatmap dos valores de entropia de metilação (ME) por gene/amostra.

De modo semelhante às análises de DMRs, a avaliação da entropia de metilação em nossas amostras permitirá identificar padrões de variabilidade epigenética que podem estar associados tanto ao tipo de tecido quanto ao doador. A entropia reflete o grau de heterogeneidade

nas modificações das citosinas, sendo, portanto, uma medida sensível para detectar alterações sutis na regulação epigenética. Com essa abordagem, poderemos investigar diferenças na estabilidade epigenética entre tecidos do mesmo indivíduo, bem como variações interindividuais com potencial impacto na expressão de genes relacionados à pigmentação. Além disso, a entropia de metilação pode ser influenciada por fatores como o envelhecimento, permitindo a exploração de possíveis efeitos da idade sobre a plasticidade epigenética em genes associados à pigmentação.

5 CONCLUSÃO

Neste trabalho, estabelecemos um conjunto de pipelines integrados para coleta de amostras, extração de DNA, sequenciamento de genomas completos com a tecnologia Oxford Nanopore e análises bioinformáticas específicas. Todas as etapas foram desenhadas com foco na aplicação em genética forense, desde a geração dos dados até a interpretação de modificações epigenéticas associadas a fenótipos. A abordagem proposta demonstrou-se viável e robusta para caracterização genômica e epigenômica, contribuindo para o avanço de estratégias preditivas baseadas em DNA no contexto forense.

REFERÊNCIAS

Aliferi, A., Sundaram, S., Ballard, D., Freire-Aradas, A., Phillips, C., Lareu, M. V., & Court, D. S. (2022). Combining current knowledge on DNA methylation-based age estimation towards the development of a superior forensic DNA intelligence tool. *Forensic Science International: Genetics*, *57*, 102637. https://doi.org/10.1016/j.fsigen.2021.102637

Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., de Cesare, M., Oikkonen, L. E., Parkes, D., Freeman, C., Dhalla, F., Patel, S. Y., Popitsch, N., Ip, C. L. C., Roberts, H. E., Salatino, S., Lockstone, H., Lunter, G., Taylor, J. C., Buck, D., Simpson, M. A., & Donnelly, P. (2019). Sequencing of human genomes with nanopore technology. *Nature Communications*, *10*(1), 1869. https://doi.org/10.1038/s41467-019-09637-5

Breslin, K., Wills, B., Ralf, A., Ventayol Garcia, M., Kukla-Bartoszek, M., Pospiech, E., Freire-Aradas, A., Xavier, C., Ingold, S., de La Puente, M., van der Gaag, K. J., Herrick, N., Haas, C., Parson, W., Phillips, C., Sijen, T., Branicki, W., Walsh, S., & Kayser, M. (2019). HIrisPlex-S system for eye, hair, and skin color prediction from DNA: Massively parallel sequencing solutions for two common forensically used platforms. *Forensic Science International: Genetics*, *43*, 102152. https://doi.org/10.1016/j.fsigen.2019.102152

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Paul Flicek, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., … Xiao, C. (2022).

High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Carratto, T. M. T., Marcorin, L., Debortoli, G., Silva, G. V., Fracasso, N. C. A., Oliveira, M. L. G., Pereira, A. L. E., Silva, A. B. C., Donadi, E. A., Simões, A. L., Castelli, E. C., Norton, H. L., Parra, E. J., & Mendes-Junior, C. T. (2019). Evaluation of the HIrisPlex-S system in a Brazilian population sample. *Forensic Science International: Genetics Supplement Series*, *7*(1), 794–796. https://doi.org/10.1016/j.fsigss.2019.10.180

Carratto, T. M. T., Marcorin, L., do Valle-Silva, G., de Oliveira, M. L. G., Donadi, E. A., Simões, A. L., Castelli, E. C., & Mendes-Junior, C. T. (2021). Prediction of eye and hair pigmentation phenotypes using the HIrisPlex system in a Brazilian admixed population sample. *International Journal of Legal Medicine*, *135*(4), 1329–1339. https://doi.org/10.1007/s00414-021-02554-7

Carratto, T. M. T., Moraes, V. M. S., Recalde, T. S. F., Oliveira, M. L. G. de, & Teixeira Mendes-Junior, C. (2022). Applications of massively parallel sequencing in forensic genetics. *Genetics and Molecular Biology*, *45*(3 suppl 1). https://doi.org/10.1590/1678-4685-gmb-2022-0077

Chaitanya, L., Breslin, K., Zuñiga, S., Wirken, L., Pośpiech, E., Kukla-Bartoszek, M., Sijen, T., Knijff, P. de, Liu, F., Branicki, W., Kayser, M., & Walsh, S. (2018). The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Science International: Genetics*, *35*, 123–135. https://doi.org/10.1016/j.fsigen.2018.04.004

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Cornelis, S., Gansemans, Y., Deleye, L., Deforce, D., & Van Nieuwerburgh, F. (2017). Forensic SNP Genotyping using Nanopore MinION Sequencing. *Scientific Reports*, *7*(1), 41759. https://doi.org/10.1038/srep41759

Cornelis, S., Gansemans, Y., Vander Plaetsen, A.-S., Weymaere, J., Willems, S., Deforce, D., & Van Nieuwerburgh, F. (2019). Forensic tri-allelic SNP genotyping using nanopore sequencing. *Forensic Science International: Genetics*, *38*, 204–210. https://doi.org/10.1016/j.fsigen.2018.11.012

De Coster, W., & Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*, *39*(5). https://doi.org/10.1093/bioinformatics/btad311

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PLoS ONE*, *16*(10 October). https://doi.org/10.1371/journal.pone.0257521

Ferreira, M. R., Carratto, T. M. T., Frontanilla, T. S., Bonadio, R. S., Jain, M., de Oliveira, S. F., Castelli, E. C., & Mendes-Junior, C. T. (2025). Advances in forensic genetics: Exploring the potential of long read sequencing. *Forensic Science International: Genetics*, *74*, 103156. https://doi.org/10.1016/j.fsigen.2024.103156

Freire-Aradas, A., Girón-Santamaría, L., Mosquera-Miguel, A., Ambroa-Conde, A., Phillips, C., Casares de Cal, M., Gómez-Tato, A., Álvarez-Dios, J., Pospiech, E., Aliferi, A., Syndercombe Court, D., Branicki, W., & Lareu, M. V. (2022). A common epigenetic clock from childhood to old age. *Forensic Science International: Genetics*, *60*, 102743. https://doi.org/10.1016/j.fsigen.2022.102743

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. https://arxiv.org/abs/1207.3907

Gershman, A., Sauria, M. E. G., Guitart, X., Vollger, M. R., Hook, P. W., Hoyt, S. J., Jain, M., Shumate, A., Razaghi, R., Koren, S., Altemose, N., Caldas, G. V., Logsdon, G. A., Rhie, A., Eichler, E. E., Schatz, M. C., O'Neill, R. J., Phillippy, A. M., Miga, K. H., & Timp, W. (2022). Epigenetic patterns in a complete human genome. *Science*, *376*(6588). https://doi.org/10.1126/science.abj5089

Jain, M. (2021). *From kilobases to "whales": a short history of ultra-long reads and high-throughput genome sequencing*. https://nanoporetech.com/about-us/news/blog-kilobases-whales-short-history-ultra-long-reads-and-high-throughput-genome

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

Kolmogorov, M., Billingsley, K. J., Mastoras, M., Meredith, M., Monlong, J., Lorig-Roach, R., Asri, M., Jerez, P. A., Malik, L., Dewan, R., Reed, X., Genner, R. M., Daida, K., Behera, S., Shafin, K., Pesout, T., Prabakaran, J., Carnevali, P., (NABEC), N. A. B. E. C., … Paten, B. (2023). Scalable Nanopore sequencing of human genomes provides a comprehensive view

of haplotype-resolved variation and methylation. *BioRxiv*. https://doi.org/10.1101/2023.01.12.523790

Kono, N., & Arakawa, K. (2019). Nanopore sequencing: Review of potential applications in functional genomics. *Development, Growth & Differentiation*, *61*(5), 316–326. https://doi.org/10.1111/dgd.12608

Lee, D., Koo, B., Yang, J., & Kim, S. (2023). Metheor: Ultrafast DNA methylation heterogeneity calculation from bisulfite read alignments. *PLOS Computational Biology*, *19*(3), e1010946. https://doi.org/10.1371/journal.pcbi.1010946

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lin, B., Hui, J., & Mao, H. (2021). Nanopore Technology and Its Applications in Gene Sequencing. *Biosensors*, *11*(7), 214. https://doi.org/10.3390/bios11070214

Lin, J.-H., Chen, L.-C., Yu, S.-C., & Huang, Y.-T. (2022). LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. *Bioinformatics*, *38*(7), 1816–1822. https://doi.org/10.1093/bioinformatics/btac058

Lin, M. F., Rodeh, O., Penn, J., Bai, X., Reid, J. G., Krasheninina, O., & Salerno, W. J. (2018). GLnexus: joint variant calling for large cohort sequencing. *BioRxiv*, 343970. https://doi.org/10.1101/343970

Ling, X., Wang, C., Li, L., Pan, L., Huang, C., Zhang, C., Huang, Y., Qiu, Y., Lin, F., & Huang, Y. (2023). Third-generation sequencing for genetic disease. *Clinica Chimica Acta*, *551*, 117624. https://doi.org/10.1016/j.cca.2023.117624

Liu, Y., Rosikiewicz, W., Pan, Z., Jillette, N., Wang, P., Taghbalout, A., Foox, J., Mason, C., Carroll, M., Cheng, A., & Li, S. (2021). DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02510-z

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Nanoporetech. (2017). *1D squared kit available in the store: boost accuracy, simple prep*.

Nanoporetech. (2023a). *Dorado*.

Nanoporetech. (2023b). *Medaka*. https://github.com/nanoporetech/medaka

Nanoporetech. (2023c). *Modkit*. https://github.com/nanoporetech/modkit

Nanoporetech. (2023d). *Remora*.

Ni, P., Nie, F., Zhong, Z., Xu, J., Huang, N., Zhang, J., Zhao, H., Zou, Y., Huang, Y., Li, J., Xiao, C.-L., Luo, F., & Wang, J. (2023). DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nature Communications*, *14*(1), 4054. https://doi.org/10.1038/s41467-023-39784-9

Nie, F., Ni, P., Huang, N., Zhang, J., Wang, Z., Xiao, C., Luo, F., & Wang, J. (2024). De novo diploid genome assembly using long noisy reads. *Nature Communications*, *15*(1), 2964. https://doi.org/10.1038/s41467-024-47349-7

Pośpiech, E., Wojas-Pelc, A., Walsh, S., Liu, F., Maeda, H., Ishikawa, T., Skowron, M., Kayser, M., & Branicki, W. (2014). The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Science International: Genetics*, *11*, 64–72. https://doi.org/10.1016/j.fsigen.2014.01.012

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ren, Z.-L., Zhang, J.-R., Zhang, X.-M., Liu, X., Lin, Y.-F., Bai, H., Wang, M.-C., Cheng, F., Liu, J.-D., Li, P., Kong, L., Bo, X.-C., Wang, S.-Q., Ni, M., & Yan, J.-W. (2021). Forensic nanopore sequencing of STRs and SNPs using Verogen's ForenSeq DNA Signature Prep Kit and MinION. *International Journal of Legal Medicine*, *135*(5), 1685–1693. https://doi.org/10.1007/s00414-021-02604-0

Ruiz, Y., Phillips, C., Gomez-Tato, A., Alvarez-Dios, J., Casares de Cal, M., Cruz, R., Maroñas, O., Söchtig, J., Fondevila, M., Rodriguez-Cid, M. J., Carracedo, Á., & Lareu, M. V. (2013). Further development of forensic eye color predictive tests. *Forensic Science International: Genetics*, *7*(1), 28–40. https://doi.org/10.1016/j.fsigen.2012.05.009

Searle, B., Müller, M., Carell, T., & Kellett, A. (2023). Third-Generation Sequencing of Epigenetic DNA. *Angewandte Chemie International Edition*, *62*(14). https://doi.org/10.1002/anie.202215704

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook,

J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*(9), 1044–1053. https://doi.org/10.1038/s41587-020-0503-6

Tytgat, O., Škevin, S., Deforce, D., & Van Nieuwerburgh, F. (2022). Nanopore sequencing of a forensic combined STR and SNP multiplex. *Forensic Science International: Genetics*, *56*, 102621. https://doi.org/10.1016/j.fsigen.2021.102621

Vidaki, A., & Kayser, M. (2018). Recent progress, methods and perspectives in forensic epigenetics. *Forensic Science International: Genetics*, *37*, 180–195. https://doi.org/10.1016/j.fsigen.2018.08.008

Walsh, S., Liu, F., Ballantyne, K. N., van Oven, M., Lao, O., & Kayser, M. (2011). IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Science International: Genetics*, *5*(3), 170–180. https://doi.org/10.1016/j.fsigen.2010.02.004

Wan, Y. K., Hendra, C., Pratanwanich, P. N., & Göke, J. (2022). Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data. *Trends in Genetics*, *38*(3), 246–257. https://doi.org/10.1016/j.tig.2021.09.001

Wang, J., Yang, L., Cheng, A., Tham, C.-Y., Tan, W., Darmawan, J., de Sessions, P. F., & Wan, Y. (2024). Direct RNA sequencing coupled with adaptive sampling enriches RNAs of interest in the transcriptome. *Nature Communications*, *15*(1), 481. https://doi.org/10.1038/s41467-023-44656-3

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*, *39*(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Wescoe, Z. L., Schreiber, J., & Akeson, M. (2014). Nanopores Discriminate among Five C5-Cytosine Variants in DNA. *Journal of the American Chemical Society*, *136*(47), 16582–16587. https://doi.org/10.1021/ja508527b

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., Snutch, T. P., Loman, N., Paten, B., Loose, M., … Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods*, *16*(12), 1297–1305. https://doi.org/10.1038/s41592-019-0617-2

Woźniak, A., Heidegger, A., Piniewska-Róg, D., Pośpiech, E., Xavier, C., Pisarek, A., Kartasińska, E., Boroń, M., Freire-Aradas, A., Wojtas, M., de la Puente, M., Niederstätter, H., Płoski, R., Spólnicka, M., Kayser, M., Phillips, C., Parson, W., & Branicki, W. (2021). Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones. *Aging*, *13*(5), 6459–6484. https://doi.org/10.18632/aging.202783

Xie, H., Wang, M., de Andrade, A., Bonaldo, M. de F., Galat, V., Arndt, K., Rajaram, V., Goldman, S., Tomita, T., & Soares, M. B. (2011). Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Research*, *39*(10), 4099–4108. https://doi.org/10.1093/nar/gkr017

Yun, T., Li, H., Chang, P.-C., Lin, M. F., Carroll, A., & McLean, C. Y. (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, *36*(24), 5582–5589. https://doi.org/10.1093/bioinformatics/btaa1081

Zheng, Z., Li, S., Su, J., Leung, A. W.-S., Lam, T.-W., & Luo, R. (2022). Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, *2*(12), 797–803. https://doi.org/10.1038/s43588-022-00387-x

ANEXO

ORIENTAÇÕES/COORIENTAÇÕES

Julia Ferreira de Moraes, Coorientação de Iniciação cientifica concluída em 2023 (Colaboração com o Prof. Dr. Willian Fernando Zambuzzi do Departamento de Ciências Químicas e Biológicas do Instituto de Biociências de Botucatu - UNESP).

Julia Ferreira de Moraes, Coorientação de Mestrado em andamento em Biotecnologia com previsão de término em março de 2026 (Colaboração com o Prof. Dr. Willian Fernando Zambuzzi do Departamento de Ciências Químicas e Biológicas do Instituto de Biociências de Botucatu - UNESP).

Matheus Amaral Silva, Coorientação de Mestrado em andamento em Biotecnologia concluída em abril de 2025 (Colaboração com o Prof. Dr. Willian Fernando Zambuzzi do Departamento de Ciências Químicas e Biológicas do Instituto de Biociências de Botucatu - UNESP).

DISCIPLINAS MINISTRADAS COM A RESPECTIVA CARGA HORÁRIA E/OU REALIZAÇÃO DE ATIVIDADES DE EXTENSÃO, COMENTÁRIOS DO SUPERVISOR)

Participação na disciplina “Tópicos em Genética e Biologia Molecular Forense”. Carga horária ministrada de 1 hora.

PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS NO PERÍODO

Eventos científicos como organizador

I WORKSHOP DE BIOINFORMÁTICA APLICADA À GENÉTICA FORENSE: ANÁLISE DE DADOS DE SEQUENCIAMENTO DE SEGUNDA E TERCEIRA GERAÇÃO. (Workshop virtual)

II WORKSHOP DE BIOINFORMÁTICA APLICADA À GENÉTICA FORENSE: ANÁLISE DE DADOS DE SEQUENCIAMENTO DE SEGUNDA E TERCEIRA GERAÇÃO. (Workshop virtual)

Eventos científicos como congressista

XVIII Congresso de Física Aplicada à Medicina. Organizado pelos alunos do curso de Física Médica da Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Campus de Botucatu.

IX WORKBIOTECH: Workshop da Pós-graduação em Biotecnologia, da Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Campus de Botucatu.

FIGG Training Brazil, realizado pela QIAGEN de maneira virtual, nos dias 17 a 20 de junho de 2024.

XIII Congresso de Biociências. Organizado pelos alunos dos cursos da Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Campus de Botucatu.

Eventos científicos como palestrante

I WEBB – WORKSHOP DE ENGENHARIA DE BIOPROCESSOS E BIOTECNOLOGIA. Workshop presencial organizado pelos alunos do curso de Engenharia de Bioprocessos e Biotecnologia da Faculdade de Ciências Agronômicas (FCA-UNESP) de Botucatu. Palestra: "*Biotecnologia Forense: Bioinformática aplicada na identificação humana por DNA*"

XVIII Congresso de Física Aplicada à Medicina. Organizado pelos alunos do curso de Física Médica da Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Campus de Botucatu. Palestra: "*Bioinformática: Perspectivas para Físicos Médicos*"

XIII Congresso de Biociências. Organizado pelos alunos dos cursos da Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências, Campus de Botucatu. Minicurso: "*Desvendando os segredos da visualização de dados: ggplot2 para iniciantes*"

Eventos científicos como avaliador

XVIII Congresso de Física Aplicada à Medicina.

XXXVI Congresso de Iniciação Científica da Unesp - IB/Botucatu.

Participações em bancas

MEMBRO TITULAR da Comissão Examinadora da DEFESA DE DISSERTAÇÃO de **RENATO MATTOS**, discente regular do Programa de Pós-Graduação em Biologia Geral e Aplicada, Curso de Mestrado Acadêmico, cujo trabalho se intitula *Caracterização funcional do miR-184 na próstata de ratos submetidos a má nutrição materna. Potencial participação na origem desenvolvimentista do câncer de próstata*.

MEMBRO TITULAR da Comissão Examinadora do EXAME GERAL DE QUALIFICAÇÃO de **ANA BEATRIZ RODRIGUES**, discente regular do Programa de Pós-Graduação em Ciências Biológicas (Genética), Curso de Mestrado Acadêmico, cujo trabalho se

intitula *Análise do Perfil Transcricional de RNAs Longos Não Codificantes em Pacientes Pediátricos com Infecção por Vírus Sincicial Respiratório*.

MEMBRO TITULAR da Comissão Examinadora da DEFESA DE DISSERTAÇÃO de **MARIA GABRIELA JACHETO CARRA**, discente regular do Programa de Pós-Graduação em Biologia Geral e Aplicada, Curso de Mestrado Acadêmico, cujo trabalho se intitula *Síntese e caracterização de revestimento de fosfato de cálcio dopado com magnésio sobre substrato de titânio*.

MEMBRO TITULAR da Comissão Examinadora do EXAME GERAL DE QUALIFICAÇÃO de **KARINA FIGUEIREDO HENRIQUES**, discente regular do Programa de Programa de Pós-Graduação em Biologia Animal, Curso de Mestrado Acadêmico, cujo trabalho se intitula *Pesquisa de variantes gênicas raras por sequenciamento completo de exoma em pacientes com deficiência intelectual sindrômica e cariótipo 46,XX*.

MEMBRO TITULAR da banca de avaliação do Trabalho de Conclusão de Curso – TCC do(a) aluno(a) **CARLOS HENRIQUE SILVA SIMÕES**, do curso de graduação de Engenharia de Bioprocessos e Biotecnologia desta Unidade Universitária, defendido na data 30/10/2024, com o título *Avaliação da expressão de lncRNAs associada ao envelhecimento na próstata ventral de ratos*.

LISTA DE PUBLICAÇÕES NO PERÍODO

KIR-MAPPER: A TOOLKIT FOR KILLER-CELL IMMUNOGLOBULIN-LIKE RECEPTOR (KIR) GENOTYPING FROM SHORT-READ SECOND-GENERATION SEQUENCING DATA. Castelli, E.C., Pereira, R.N., Paes, G.S., Andrade, H.S., Ferreira, M.R., de Freitas Santos, Í.S., Vince, N., Pollock, N.R., Norman, P.J. and Meyer, D. (2025). HLA, 105: e70092. https://doi.org/10.1111/tan.70092

ADVANCES IN FORENSIC GENETICS: EXPLORING THE POTENTIAL OF LONG READ SEQUENCING. Marcel Rodrigues Ferreira, Thássia Mayra Telles Carratto, Tamara Soledad Frontanilla, Raphael Severino Bonadio, Miten Jain, Silviene Fabiana de Oliveira, Erick C. Castelli, Celso Teixeira Mendes-Junior, Forensic Science International: Genetics, Volume 74, 2025, 103156, ISSN 1872-4973, https://doi.org/10.1016/j.fsigen.2024.103156.

EFFECTS OF EMPAGLIFLOZIN ON MYOCARDIAL TRANSCRIPTOME IN RATS WITH AORTIC STENOSIS-INDUCED HEART FAILURE. L Urbano Pagan, M Gatto, M R Ferreira, M J Gomes, J P G Oliveira, G A F Mota, F C Damatto, L M Souza, A C C Santos, E A Rodrigues, P A Borim, D H S Campos, M P Okoshi, K Okoshi, European Heart

Journal, Volume 45, Issue Supplement_1, October 2024, ehae666.1829, https://doi.org/10.1093/eurheartj/ehae666.1829

IMPACT OF MATERNAL PROTEIN RESTRICTION ON THE PROTEOMIC LANDSCAPE OF MALE RAT LUNGS ACROSS THE LIFESPAN. Matheus Naia Fioretto, Flávia Alessandra Maciel, Luísa Annibal Barata, Isabelle Tenori Ribeiro, Carolina Beatriz Pinheiro Basso, Marcel Rodrigues Ferreira, Sérgio Alexandre Alcantara dos Santos, Renato Mattos, Hecttor Sebastian Baptista, Luiz Marcos Frediane Portela, Pedro Magalhães Padilha, Sérgio Luis Felisbino, Wellerson Rodrigo Scarano, Elena Zambrano, Luis Antonio Justulin, Molecular and Cellular Endocrinology, Volume 592, 2024, 112348, ISSN 0303-7207, https://doi.org/10.1016/j.mce.2024.112348.

VASCULAR SMOOTH MUSCLE CELL-DERIVED EXOSOMES PROMOTE OSTEOBLAST-TO-OSTEOCYTE TRANSITION VIA B-CATENIN SIGNALING. Célio J.C. Fernandes, Rodrigo A. Silva, Marcel R. Ferreira, Gwenny M. Fuhler, Maikel P. Peppelenbosch, Bram CJ. van der Eerden, Willian F. Zambuzzi, Experimental Cell Research, Volume 442, Issue 1, 2024, 114211, ISSN 0014-4827, https://doi.org/10.1016/j.yexcr.2024.114211.

VASCULAR SMOOTH MUSCLE CELLS EXHIBIT ELEVATED HYPOXIA-INDUCIBLE FACTOR-1A EXPRESSION IN HUMAN BLOOD VESSEL ORGANOIDS, INFLUENCING OSTEOGENIC PERFORMANCE. Geórgia da Silva Feltran, Rodrigo Augusto da Silva, Célio Junior da Costa Fernandes, Marcel Rodrigues Ferreira, Sérgio Alexandre Alcântara dos Santos, Luis Antônio Justulin Junior, Liliana del Valle Sosa, Willian Fernando Zambuzzi, Experimental Cell Research, Volume 440, Issue 2, 2024, 114136, ISSN 0014-4827, https://doi.org/10.1016/j.yexcr.2024.114136.

ADORA2A DOWNREGULATION PROMOTES CAFFEINE NEUROPROTECTIVE EFFECT AGAINST LPS-INDUCED NEUROINFLAMMATION IN THE HIPPOCAMPUS. Paula Lemes dos Santos Sanna, Liebert Bernardes Carvalho, Camila Cristina dos Santos Afonso, Kassia de Carvalho, Rogério Aires, Jennyffer Souza, Marcel Rodrigues Ferreira, Alexander Birbrair, Maria Martha Bernardi, Alexandra Latini, Rodrigo A. Foganholi da Silva, Brain Research, Volume 1833, 2024, 148866, ISSN 0006-8993, https://doi.org/10.1016/j.brainres.2024.148866.
(https://www.sciencedirect.com/science/article/pii/S0006899324001203)

THE MULTIFARIOUS FUNCTIONS OF LEUKOTRIENES IN BONE METABOLISM. Flávia Amadeu de Oliveira, Cintia K. Tokuhara, Vimal Veeriah, João Paulo Domezi, Mariana R. Santesso, Tania M. Cestari, Talita M.O. Ventura, Adriana A. Matos, Thiago Dionísio, Marcel R. Ferreira, Rafael C. Ortiz, Marco A.H. Duarte, Marília A.R. Buzalaf, José B.

Ponce, Carlos A. Sorgi, Lucia H. Faccioli, Camila Peres Buzalaf, Rodrigo Cardoso de Oliveira, Journal of Bone and Mineral Research, Volume 38, Issue 8, 1 August 2023, Pages 1135–1153, https://doi.org/10.1002/jbmr.4867

OSTEOGENIC DIFFERENTIATION AND RECONSTRUCTION OF MANDIBLE DEFECTS USING A NOVEL RESORBABLE MEMBRANE: AN IN VITRO AND IN VIVO EXPERIMENTAL STUDY. Edmara T. P. Bergamo DDS, MSci, PhD, Ísis de Fátima Balderrama DDS, MSci, PhD, Marcel Rodrigues Ferreira MSci, PhD, Robert Spielman JD, Blaire V. Slavin BSci, Andrea Torroni MD, PhD, Nick Tovar PhD, DDS, Vasudev V. Nayak MSci, PhD, Benjamin R. Slavin MD, Paulo G. Coelho MD, DDS, PhD, MBA, Lukasz Witek MSci, PhD. J Biomed Mater Res. 2023; 111(11): 1966-1978. doi:10.1002/jbm.b.35299

COMBINATION OF IN SILICO AND CELL CULTURE STRATEGIES TO PREDICT BIOMATERIAL PERFORMANCE: EFFECTS OF SINTERING TEMPERATURE ON THE BIOLOGICAL PROPERTIES OF HYDROXYAPATITE. Gerson S. de Almeida, Marcel R. Ferreira, Célio C. Fernandes Jr, Carlos A. O. de Biagi Jr, Wilson Araújo Silva Jr, Elidiane C. Rangel, Paulo N. Lisboa-Filho, Willian F. Zambuzzi. J Biomed Mater Res. 2024; 112(2):e35389. doi:10.1002/jbm.b.35389

DEVELOPMENT OF COBALT (CO)-DOPED MONETITES FOR BONE REGENERATION. Gerson Santos de Almeida, Marcel Rodrigues Ferreira, Célio Junior da Costa Fernandes, Luísa Camilo Suter, Maria Gabriela Jacheto Carra, Diego Rafael Nespeque Correa, Elidiane Cipriano Rangel, Margarida Juri Saeki, Willian Fernando Zambuzzi. J Biomed Mater Res. 2024; 112(1):e35319. doi:10.1002/jbm.b.35319

MICRORNA BIOGENESIS MACHINERY ACTIVATION AND LNCRNA AND REST OVEREXPRESSION AS NEUROPROTECTIVE RESPONSES TO FIGHT INFLAMMATION IN THE HIPPOCAMPUS. Liebert Bernardes Carvalho, Paula Lemes dos Santos Sanna, Camila Cristina dos Santos Afonso, Eduardo F. Bondan, Geórgia da Silva Feltran, Marcel Rodrigues Ferreira, Alexander Birbrair, Denise Carleto Andia, Alexandra Latini, Rodrigo A. Foganholi da Silva, Journal of Neuroimmunology, Volume 382, 2023, 578149, ISSN 0165-5728, https://doi.org/10.1016/j.jneuroim.2023.578149.

CYCLOPAMINE TARGETING HEDGEHOG MODULATES NUCLEAR CONTROL OF THE OSTEOBLAST ACTIVITY. Célio J. da Costa Fernandes, Marcel Rodrigues Ferreira, Willian Fernando Zambuzzi, Cells & Development, Volume 174, 2023, 203836, ISSN 2667-2901, https://doi.org/10.1016/j.cdev.2023.203836.