

Universidade de São Paulo  
Instituto de Matemática e Estatística  
Programa Interunidades de Pós-graduação em Bioinformática

**Mapeamento e estudo de associação de Variações  
no Número de Cópias de DNA (CNVs) e de Regiões  
com Perda de Heterozigozidade (LOH) com  
Fenótipos Complexos**

Camila Alves de Souza

Orientação: Prof<sup>a</sup> Dra. Júlia Maria Pavan Soler

Dissertação de Mestrado ao  
Instituto de Matemática e Estatística da Universidade de São Paulo  
para obtenção do título de Mestre em Bioinformática

São Paulo  
2022

**Mapeamento e estudo de associação de Variações no  
Número de Cópias de DNA (CNVs) e de Regiões com  
Perda de Heterozigozidade (LOH) com Fenótipos  
Complexos**

Dissertação de Mestrado ao  
Instituto de Matemática e Estatística  
da Universidade de São Paulo  
para obtenção do título de  
Mestre em Bioinformática.

São Paulo  
2022

# Agradecimentos

Agradeço aos meus pais Eliana e Cosme, meu marido Denis e minha irmã Caroline, pelo apoio, compreensão, carinho e não me deixarem nunca desistir dos meus objetivos.

Agradeço a minha orientadora Professora Júlia, pelos ensinamentos, toda a compreensão e carinho ao longo de toda a minha jornada.

Agradeço aos meus amigos de trabalho do setor de Bioinformática da Dasa Genômica: Juliana, Marcelo, Felipe, Matheus, Luciane e Kroll pelo apoio e a amizade. Em especial agradeço ao Rodrigo, que me deu a oportunidade de aprender e sempre teve muito carinho e compreensão em ensinar as atividades da área de bioinformática e disponibilizou o tempo para que eu possa desenvolver as atividades acadêmicas paralelamente as atividades corporativas.

Agradeço com muito carinho a Jaqueline que me ajudou com o desenvolvimento das atividades, trocando experiências e aprendendo juntas com as diferentes habilidades que temos. Também, agradeço a Professora Regina e a toda equipe do projeto ISA-Nutrição pelo carinho e pela oportunidade de me desenvolver e aprender com os dados concedidos. E agradeço a Secretaria Municipal da Saúde de São Paulo, FAPESP e CNPq pelo apoio financeiro ao projeto ISA-Capital e ISA-Nutrição.

# Resumo

SOUZA, C.A. **Mapeamento e estudo de associação de Variações no Número de Cópias de DNA (CNVs) e de Regiões com Perda de Heterozigozidade (LOH) com Fenótipos Complexos.** Mestrado em Bioinformática – Instituto de Matemática e Estatística, Universidade de São Paulo.

Com o avanço das tecnologias citogenômicas, tornou-se possível a detecção confiável de diversas alterações genômicas e a determinação da ancestralidade dos indivíduos, uma importante ferramenta de entendimento da origem e da história das populações, bem como, fonte de decisões no contexto de saúde pública. Dentre os diferentes tipos de variação genética, existem as *Copy Number Variation* (CNV) que são ganhos ou perdas de segmentos de DNA. Diversos trabalhos associam as CNVs a diferentes tipos de fenótipos envolvidos com doenças monogênicas bem como multifatoriais. Outra variação genética importante é a *Loss of heterozygosity* (LOH), uma instabilidade cromossômica caracterizada pelo desequilíbrio alélico, onde uma região genômica heterozigota se torna homozigota através da perda de um dos alelos. As regiões homozigotas podem representar populações, determinar nível de consanguinidade e diversos estudos correlacionam as regiões LOH como uma das principais causas de tumorigênese. Neste trabalho, tendo em vista a importância da determinação da ancestralidade dos indivíduos e do impacto das variantes CNVs e das LOHs ao contexto da saúde humana, o objetivo é construir bancos de dados de ancestralidade, CNVs e LOH a partir dos dados do projeto ISA-Nutrição, genotipados através da plataforma *Axiom 2.0* (*Affymetrix, Thermo Fisher*). Tal informação permitirá correlacionar variações moleculares a fenótipos que sejam relevantes ao grupo de estudo, contribuindo assim para o melhor entendimento do perfil do brasileiro. O processamento dos dados envolvido neste trabalho não é tarefa simples pois exige a execução de diferentes etapas de processamento utilizando programas como PLINK, além de programas proprietários disponibilizados pela *Affymetrix* e bibliotecas das linguagens R e Python. Neste sentido, uma contribuição do presente trabalho é propor um processamento customizado para análise de ancestralidade e identificação de regiões de CNVs e de LOH a partir de dados de *SNParray*. **Resultados:** Como esperado, a amostra ISA-Nutrição permitiu caracterizar a alta miscigenação da população brasileira e alguns indivíduos apresentaram diferenças entre a raça autodeclarada e as proporções ancestrais das populações em comparação com o projeto 1000 genomas. Foram detectadas 34.824 CNVs para as 841 amostras do estudo, variando entre 49.3kb (CN=0) a 33.5Mb (CN=3). As LOHs tiveram média de 1,5% nos cromossomos autossômicos, foram detectadas 26.558 nas 841 amostras e variaram entre 1Mb e 108Mb.

**Palavras-chave:** Ancestralidade, *Copy Number Variation*, *Loss of heterozygosity*.

# Abstract

SOUZA, C.A. **Mapping and study of association of DNA Copy Number Variations (CNVs) and Loss of Heterozygosity Regions (LOH) with Complex Phenotypes.** Master's in bioinformatics – Institute of Mathematics and Statistics, University of São Paulo.

With the advancement of cytogenomic technologies, it became possible to reliably detect several genomic alterations and determine the ancestry of individuals, an important tool for understanding the origin and history of populations, as well as a source of decisions in the context of public health. Among the different types of genetic variation, there are Copy Number Variation (CNV) which are gains or losses of DNA segments. Several studies associate CNVs with different types of phenotypes involved with monogenic as well as multifactorial diseases. Another important genetic variation is Loss of heterozygosity (LOH), a chromosomal instability characterized by allelic imbalance, where a heterozygous genomic region becomes homozygous through the loss of one of the alleles. Homozygous regions can represent populations, determine level of consanguinity, and several studies have correlated LOH regions as one of the main causes of tumorigenesis. In this work, in view of the importance of determining the ancestry of individuals and the impact of CNV and LOH variants in the context of human health, the objective is to build ancestry databases, CNVs and LOH from data from the ISA- Nutrition, genotyped using the Axiom 2.0 platform (Affymetrix, Thermo Fisher). Such information will allow the correlation of molecular variations to phenotypes that are relevant to the study group, thus contributing to a better understanding of the Brazilian profile. The data processing involved in this work is not a simple task as it requires the execution of different pipelines that use different software such as PLINK, and proprietary software provided by Affymetrix and R and Python language libraries. In this sense, a contribution of the present work is to propose a customized pipeline for ancestry analysis and identification of CNV and LOH regions from SNParray data. **Results:** As expected, ISA-Nutrition data showed the high miscegenation of the Brazilian population, and some individuals show differences between self-reported race and ancestral proportions of populations compared to the 1000genomas project. In total 34,824 CNVs were detected for the 841 samples of the study, ranging from 49.3kb (CN=0) to 33.5Mb (CN=3). The LOHs had an average of 1.5% in the autosomal chromosomes, 26,558 were detected in the 841 samples, and ranged from 1Mb to 108Mb.

**Keywords:** Ancestry, Copy Number Variation, Loss of heterozygosity.

# Sumário

1. Introdução .....	12
1.1. <i>Genome Wide Association Study</i> - GWAS .....	12
1.2. Ancestralidade.....	13
1.3. Métricas de genética populacional.....	14
1.4. <i>Copy Number Variation</i> (CNV) .....	16
• CNVs em estudos de associação .....	19
1.5. <i>Loss of Heterozigosity</i> (LOH).....	19
2. Objetivos .....	21
3. Materiais e Métodos .....	22
3.1. Banco de Dados.....	22
3.2. Categorização dos participantes .....	23
• Declaração de Raça.....	23
• Local de Nascimento.....	23
3.2.1. Extração e métricas de qualidade do material genético.....	24
3.2.2. Plataforma de Genotipagem: <i>Axiom™ 2.0 Precision Medicine Research Array</i> .....	25
3.3. Genotipagem .....	30
• Algoritmo AxiomGT1 .....	30
3.4. Verificação de Qualidade dos SNPs.....	33
3.5. Gráfico <i>Manhattan</i> .....	35
3.6. Filtragem dos SNPs.....	35
Filtragem pelo <i>Apt-Genotype-Axiom</i> e conversão para formato PLINK .....	35
Filtro por Regiões Complexas .....	36
Filtro por Frequência .....	36
• Filtros por cromossomos determinados .....	38
3.7. Ancestralidade Global .....	38
• Comparação com os dados do Projeto 1000 Genomas.....	38
• <i>SNPRelate</i> .....	39
3.8. <i>Copy Number Variation</i> (CNV) e <i>Loss of Heterozygosity</i> (LOH).....	40
3.8.1. Construção da Referência.....	40
3.8.2. Determinação do BAF e LRR para chamada de CNVs e LOH .....	41
3.8.3. Chamada de CNV e LOH .....	43
4. Resultados e Discussão .....	45

4.1. Disponibilização dos Bancos de Dados Genômicos .....	45
4.1.1. Gráfico <i>Manhattan</i> .....	45
4.2. Análise de Ancestralidade Global - Análise de Componentes Principais (PCA) .....	47
4.3. Gráficos de Ancestralidade Global.....	52
4.2. <i>Copy Number Variation</i> (CNV) e <i>Loss of Heterozygosity</i> (LOH).....	61
4.2.1. Criação da Referência Personalizada .....	61
4.2.2. Chamadas de CNV e LOH .....	62
• Filtro de Qualidade .....	62
• Chamadas de CNV por cromossomo.....	63
• Comparação do Gênero das Amostras.....	63
• Tamanho das CNVs .....	64
• Tipos de CNV .....	66
• Porcentagem de LOH nos cromossomos autossômicos.....	67
• Tamanhos das LOHs .....	68
• Sobreposição das Chamadas de CNVs e LOH .....	70
4.2.3. Avaliação de possíveis fenótipos associados às regiões genômicas dos achados mais frequentes .....	70
• Identificação das regiões genômicas com a maior frequência de CNVs e LOH .....	70
4.2.4. Associação de % LOHs nos cromossomos autossômicos com o fenótipo de obesidade.....	71
5. Considerações Finais.....	74
6. Referências .....	75
Apêndice .....	80

# Lista de Figuras

- Figura 1. Descrição do Desequilíbrio de Ligação (Bush & Moore, 2012).
- Figura 2. Exemplo de CNVs. Adaptado de <https://geneticeducation.co.in/what-is-copy-number-variation-and-how-to-detect-it>.
- Figura 3. Mecanismos de formação das SVs. Adaptado (Weischenfeldt et al., 2013)
- Figura 4. Exemplo de trechos de LOH em um marcador em células normais e tumorais. Adaptado (Lin et al., 2004)
- Figura 5. Tabela com o grupo de marcadores da plataforma *Axiom PMRA*.
- Figura 6. Protocolo da seleção de amostra para genotipagem.
- Figura 7. Etapas para as melhores práticas de genotipagem de análise de fluxo de trabalho.
- Figura 8. Amostras com *QC call rate* abaixo do limite de qualidade (97%)
- Figura 9. Gráfico do *QC call rate* por DQC.
- Figura 10. *Boxplot* (Diagrama de Caixa) da métrica DQC para as 9 placas processadas.
- Figura 11. *Boxplot* (Diagrama de Caixa) da métrica *QC call rate* para as 9 placas processadas.
- Figura 12. Log-Transformação e Rotação de A e B, aplicados pelo algoritmo *AxiomGT1*.
- Figura 13. Conversão dos alelos para a nomenclatura A e B.
- Figura 14. Valor de cada genótipo para o algoritmo de genotipagem.
- Figura 15. Exemplos de marcadores *PolyHighResolution*. Fonte: *User Guide: Axiom™ Genotyping Solution Data Analysis (Thermo Fisher Scientific)*
- Figura 16. Exemplo de marcadores *NoMinorHom*. Fonte: *User Guide: Axiom™ Genotyping Solution Data Analysis (Thermo Fisher Scientific)*
- Figura 17. Exemplos de marcadores *MonoHighResolution*. Fonte: *User Guide: Axiom™ Genotyping Solution Data Analysis (Thermo Fisher Scientific)*
- Figura 18. Contagem de Indivíduos por Domicílio.
- Figura 19. Intensidades dos alelos A e B, e equação de determinação dos valores de R e  $\theta$ .
- Figura 20. Genótipos e seus respectivos valores de BAF.
- Figura 21. LRR e seus respectivos números de cópias.
- Figura 22. LRR e BAF e seus respectivos números de cópias. Adaptado de (Alkan et al., 2011)
- Figura 23. Exemplo da sobreposição das chamadas, adaptado de (Quinlan & Hall, 2010).
- Figura 24. Bancos Disponibilizados para o Projeto ISA - Nutrição
- Figura 25. Gráfico *Manhattan* dos valores p do teste HWE
- Figura 26. Na esquerda (em verde) o número de SNPs ao longo dos cromossomos antes da etapa de remoção por *pruning* e na direita (em azul) o número de SNPs para os cromossomos depois da etapa de remoção por *pruning*.
- Figura 27. Porcentagem da explicação dos 6 primeiros componentes principais com ancestralidade.



Figura 28. Na esquerda a correlação do CP1 (51,1% de explicação) e na direita a correlação do CP2 (25.4%) com os genótipos ao longo do genoma.

Figura 29. Relacionamento dos 6 primeiros CP da análise do Projeto ISA-Nutrição + Projeto 1000 genomas.

Figura 30. Porcentagem da explicação dos 6 primeiros componentes principais com ancestralidade.

Figura 31. Na esquerda a correlação do CP1 (15.2% de explicação) e na direita a correlação do CP2 (9.2%) com os genótipos ao longo do genoma.

Figura 32. Relacionamento dos 6 primeiros CP da análise do Projeto ISA-Nutrição.

Figura 33. (1) Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição pelos CP1 e CP2. (2) Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição pelos CP1 e CP2 com destaque nos 134 indivíduos relacionados, excluídos na primeira etapa do cálculo.

Figura 34. Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição categorizados por raça autodeclarada pelos CP1 e CP2. Na primeira figura com a informação das superpopulações e na segundo o destaque nas subcategorias do projeto ISA-Nutrição.

Figura 35. Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição categorizados pelo local de nascimento pelos CP1 e CP2. Na primeira figura com a informação das superpopulações e na segundo o destaque nas subcategorias do projeto ISA-Nutrição.

Figura 36. Proporções de ancestralidade Individual das amostras do Projeto ISA-Nutrição.

Figura 37. Proporções de Ancestralidade Individual de autodeclarados brancos.

Figura 38. Proporções de Ancestralidade Individual de autodeclarados pardos.

Figura 39. Proporções de Ancestralidade Individual de autodeclarados pretos.

Figura 40. Proporção média de ancestralidade por raça autodeclarada. Para cada população tivemos: ISA-Amarela=13, ISA-Branca=422, ISA-Indígena=2, ISA-NR=11, ISA-Parda=309, ISA-Preta=84 (Total 841).

Figura 41. Proporção Média de Ancestralidade por Região. Para cada população tivemos: NORDESTE=190, NORTE/CENTRO OESTE=5, ESTRANGEIRO=26, NA=12, SUDESTE=584 e SUL=24 (Total 841).

Figura 42. (1) Histograma e (2) *Boxplot* (Diagrama de Caixa) mostrando os valores de MAPD para as 841 amostras do Projeto. Ponto de corte recomendado: excluir amostras com valores acima de 0,35.

Figura 43. (1) Histograma e (2) *Boxplot* (Diagrama de Caixa) mostrando os valores de *WavinessSD* para as 841 amostras do Projeto. Ponto de corte recomendado: excluir valores acima de 0,10.

Figura 44. Distribuição das amostras do ISA-Nutrição aprovadas e reprovadas na comparação com filtros das métricas de (1) MAPD e (2) *WavinessSD*.

Figura 45. Valores mínimos de qualidade para MAPD e *WavinessSD* (1) e Valores mínimos para os filtros de chamadas de CNV (2).

Figura 46. Número de CNVs detectadas por cromossomo.

Figura 47. Comparação dos valores computados do cromossomo X (chrX) e do cromossomo Y (chrY) nas 841 amostras do estudo.

Figura 48. Categorias de tamanho das CNVs e os valores de chamadas detectadas durante o processamento.

Figura 49. Número de Cópias por Segmento Cromossômico. O algoritmo separa as deleções nas classes CN=0, CN=1, cópias normais CN=2 e CN=3 para as duplicações.

Figura 50. Categorias de tamanho das LOHs e os valores de chamadas detectadas durante o processamento.

Figura 51. Frequência do Fenótipo Obesidade nas 707 amostras do Projeto-ISA

Figura 52. Gráfico de pontos com a categoria de obesidade e o valor % de LOH nos cromossomos autossômicos.

# Lista de Tabelas

Tabela 1. Breve descrição dos principais métodos de detecção de CNVs. Adaptado (Pös et al., 2021)

Tabela 2. Descrição dos participantes do projeto ISA-Nutrição por gênero e faixa etária.

Tabela 3. Informações sobre os participantes do Projeto ISA-Nutrição

Tabela 4. Distribuição dos marcadores por classe de qualidade.

Tabela 5. Participantes do Projeto 1000 genomas utilizados na combinação

Tabela 6. Número de SNPs excluídos por filtros e os valores restantes durante o processamento.

Tabela 7. Estatísticas resumo dos 4 primeiros Componentes Principais (escores) encontrados nos dados do Projeto ISA-Nutrição em conjunto com os indivíduos do Projeto 1000 genomas.

Tabela 8. Número de SNPs excluídos por filtros e os valores restantes para o processamento da ancestralidade global dos 841 indivíduos.

Tabela 9. Estatísticas resumo dos 4 primeiros Componentes Principais encontrados nos dados do Projeto ISA-Nutrição.

Tabela 10. Distribuição dos 841 indivíduos do projeto ISA-Nutrição nas categorias de gênero autodeclaradas e identificadas pelo algoritmo.

Tabela 11. Estatísticas dos tamanhos (em pares de base – bp) das CNVs por cromossomo (chr).

Tabela 12. Frequência de cada tipo de CN (Número de Cópia) por cromossomo (chr).

Tabela 13. Estatísticas resumo da % de LOH nos cromossomos autossômicos.

Tabela 14. Estatísticas dos tamanhos (pares de base – bp) das LOHs por cromossomo (chr).

Tabela 15. Valores dos coeficientes da Regressão Logística de % de LOH e Obesidade.

# 1. Introdução

## 1.1. Genome Wide Association Study - GWAS

As informações genéticas têm ajudado na melhor compreensão de diversas doenças humanas, contribuindo para grandes e importantes mudanças nos cuidados com a saúde. Os grandes avanços tecnológicos na área genômica têm permitido interpretar e descobrir novos mecanismos para o diagnóstico e tratamento dos pacientes, salvando milhares de vidas (Khoury & Dotson, 2021). Algumas doenças humanas como a Anemia Falciforme e a Fibrose Cística são causadas por mutações em um único gene, conhecidas como doenças monogênicas. Contudo, doenças comuns como a diabetes e a hipercolesterolemia, possuem uma causa bem mais complexa, pela associação dos efeitos de vários genes em combinação com o estilo de vida e fatores ambientais. Estas doenças são conhecidas como Doenças Multifatoriais. (Cordell & Clayton, 2005)

Atualmente existem várias técnicas para o estudo das Doenças Multifatoriais, como os Estudos de Associação Genômica Ampla (GWAS - *Genome Wide Association Studies*). Os estudos baseados em GWAS têm sido largamente utilizados por combinarem a informação de vários sítios, permitindo a análise de muitos genes associados, potencializando as descobertas dos mecanismos envolvidos em vias metabólicas e genes que codificam proteínas com relações funcionais. (Uffelmann et al., 2021). Essa associação, em geral, é baseada em estudos observacionais Caso-Controle, isto é, na comparação das diferenças nas frequências alélicas ou genotípicas observadas em indivíduos afetados por um certo fenótipo (caso) em comparação com indivíduos que não são afetados (controle). A premissa é determinar os sítios cujas variações são exclusivamente encontradas nos indivíduos denominados caso, ou seja, ausentes nos indivíduos controle (Bush & Moore, 2012) e (Cordell & Clayton, 2005).

O sucesso de GWAS na descoberta ou confirmação da associação entre variantes moleculares e as doenças/fenótipos na população humana é muito influenciado pelo plano amostral e experimental do estudo, como o cuidado ao selecionar o número de amostras e de representação dos grupos caso e controle, e pelo processamento e análise de dados, de modo a evitar a influência de variáveis confundidoras ao estudo que enviesam os resultados. Sendo assim, é fundamental e indispensável a inclusão de um rígido Controle de Qualidade (CQ) para dados genotipados (Peterson et al., 2019).

Uma importante ferramenta de Controle de Qualidade dos dados importados no GWAS é a caracterização da ancestralidade dos indivíduos na pesquisa, para correta representação entre populações, tendo em vista que, para alguns sítios, as diferentes ancestrias podem produzir diferenças consideráveis nas Frequências Alélicas (*Allele Frequency - AF*) e nos padrões de

Desequilíbrio de Ligação (*Linkage Disequilibrium - LD*). Sendo assim, conhecendo e ajustando os modelos pelas diferentes populações, é possível minimizar uma das fontes de resultados falso-positivos. (Barrett & Cardon, 2006)

## 1.2. Ancestralidade

O conceito de raça é complexo e objeto de diversos estudos sociais. Evidências físicas como tonalidade da pele e alguns traços físicos são usados erroneamente para tentar classificar indivíduos em grupos com a mesma ancestralidade. A definição de raça tem grande embasamento em fatores socioculturais e ainda apresentam um grande apelo de segregação (Dóra et al., 2019).

Desde o período colonial, a região sudeste do Brasil sempre foi o local com o maior desenvolvimento dos recursos financeiros e onde a população encontrava as melhores oportunidades de emprego e moradia. Muitas populações de outras regiões brasileiras migraram para o sudeste motivadas pela esperança de melhores condições de vida, e até mesmo populações estrangeiras que ora recebiam ajuda do próprio governo brasileiro ou exílio em períodos de crise em seus países natais (IBGE, 2007).

O Brasil foi alvo de grandes ondas migratórias. A primeira entre os anos de 1500 e 1800, foi composta principalmente por africanos e europeus, que em sua maioria eram portugueses e homens. Após a abolição da atividade escrava africana, entre os anos de 1875 e 1975, ocorreu uma segunda grande onda migratória, incentivada pelo governo brasileiro com o objetivo de substituir a mão de obra de diversos campos econômicos que exploravam os recursos de escravos. Nesta onda, vieram principalmente cidadãos europeus e asiáticos, em maioria representadas pelas populações: italiana, espanhola, portuguesa, alemã, japonesa e até mesmo de árabes. A distribuição desses imigrantes não foi totalmente equilibrada e boa parte se tornou residente das regiões sudeste e sul do Brasil (IBGE, 2007).

Com o avanço da área da genética humana, o conhecimento das origens ancestrais dos indivíduos se tornou mais robusto. Além disso, muitas áreas da saúde se beneficiaram com o melhor entendimento de algumas doenças, considerando que elas apresentam diferentes comportamentos em cada indivíduo e que essas diferenças podem aparecer de forma relevante entre populações humanas. Áreas como a farmacogenômica se tornaram mais estruturadas ao conhecer ancestralidade genética, e tanto na resposta a medicamentos, bem como na medicina personalizada, este conceito desempenha um papel central (Yang et al., 2021).

Para conhecer a ancestralidade dos indivíduos são utilizados conceitos de genética populacional, como a composição genética das populações em termos das mudanças de genótipo e frequência

dos fenótipos em resposta a diversos processos como seleção natural, deriva genética e mutação (Neher & Shraiman, 2011).

### 1.3. Métricas de genética populacional

Um importante conceito de genética populacional é a lei do Equilíbrio de *Hardy-Weinberg*. Sob esta lei, assume-se que uma população infinitamente grande, sem sofrer seleção natural, mutação ou migração tenha seus genótipos e probabilidades alélicas constantes ao longo das gerações. Ao violar essa lei é indicado que as probabilidades alélicas que compõem os genótipos são diferentes do esperado sob independência, ou seja, podem ser indicativos de uma associação genética com risco de uma doença ou até mesmo um erro no estudo de genotipagem (Marees et al., 2018). Nos estudos de GWAS, essa lei é utilizada como um dos critérios para avaliar a estrutura e qualidade dos dados. Sob equilíbrio, para um loco cromossômico definido pela ocorrência dos alelos A e a, com probabilidades de ocorrência,  $P(A)=p$  e  $P(a)=(1-p)$ , respectivamente, as probabilidades genotípicas obedecem à independência alélica, sendo dadas por:

$$P(AA) = P(A)P(A) = p^2$$

$$P(Aa) = 2P(A)P(a) = 2p(1 - p)$$

$$P(aa) = P(a)P(a) = (1 - p)^2$$

O teste da hipótese de equilíbrio de *Hardy-Weinberg* pode ser formulado por diferentes estatísticas ((Puig et al., 2019); (Foulkes, 2009)), sendo comumente usado o teste Qui-quadrado para verificar se valores obtidos pelos dados do estudo (frequência observada  $O_i$  na classe genotípica  $i$ ) correspondem ao que é esperado pelo modelo teórico (frequência esperada  $E_i$  na classe genotípica  $i$ ). A seguinte estatística é calculada, a qual segue uma distribuição Qui-quadrado com  $(K-1)$  graus de liberdade

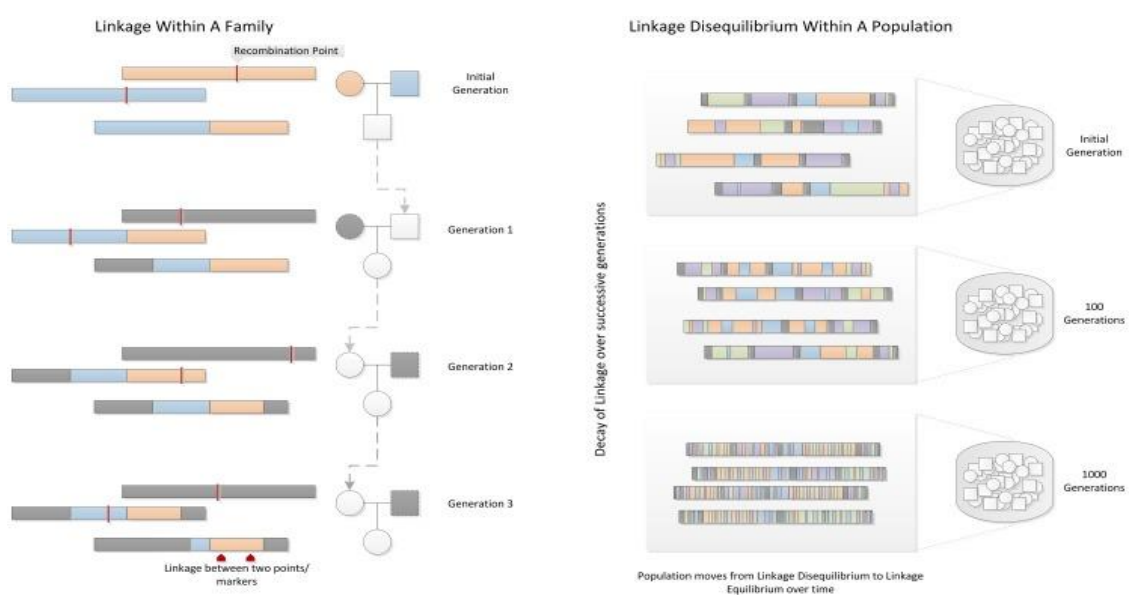
$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Outro importante conceito de genética populacional é a caracterização da Probabilidade do Alelo Menor (*Minor Allele Frequency – MAF*) em um sítio cromossômico. Sua importância está em conseguir diferenciar os alelos comuns dos alelos raros dentro da população. O excesso de alelos

comuns indica que a população pode ter sofrido um efeito de gargalo através da seleção natural ou pela homogeneidade populacional. Quando é alta a frequência de alelos raros, pode indicar a expansão da população evidenciando as taxas de mutação e fluxo gênico (Linck & Battey, 2019).

O Desequilíbrio de Ligação (LD) é outra métrica importante que mede a associação entre alelos presentes em dois ou mais sítios, diferentemente do desequilíbrio de *Hardy-Weinberg* que mede a associação entre alelos dentro do sítio e não entre sítios. Os alelos permanecem ligados em um cromossomo em vez de serem separados por eventos de recombinação durante a meiose. A extensão de LD depende de vários fatores, incluindo o tamanho da população, o número de cromossomos fundadores na população e o número de gerações para as quais a população existe. Logo, é esperado que diferentes subpopulações humanas tenham diferentes padrões de LD. Um exemplo é a observação das populações africanas, em que, como são mais antigas, apresentam menos regiões em Desequilíbrio de Ligação, tendo em vista que existe um maior acúmulo de eventos de recombinação que ocorre durante um certo tempo. Populações mais recentes, que são mais suscetíveis ao efeito fundador, têm, em média, mais regiões em LD, por terem tido menos tempo para recombinação no genoma (Bush & Moore, 2012). Uma outra fonte de LD é a migração, sob a qual um desbalanceamento na composição alélica da população pode ocorrer levando ao desvio da independência na distribuição dos alelos.

O conceito de LD descreve, matematicamente, as mudanças das variações genéticas em uma população ao longo de gerações (Figura 1). Durante os eventos de recombinações aleatórias como o *crossing over* (nas cromátides homólogas) durante a meiose, eventualmente, os alelos das populações ao longo do tempo tendem a entrar em equilíbrio de ligação, ou seja, se tornam mais diferenciados (Bush & Moore, 2012).

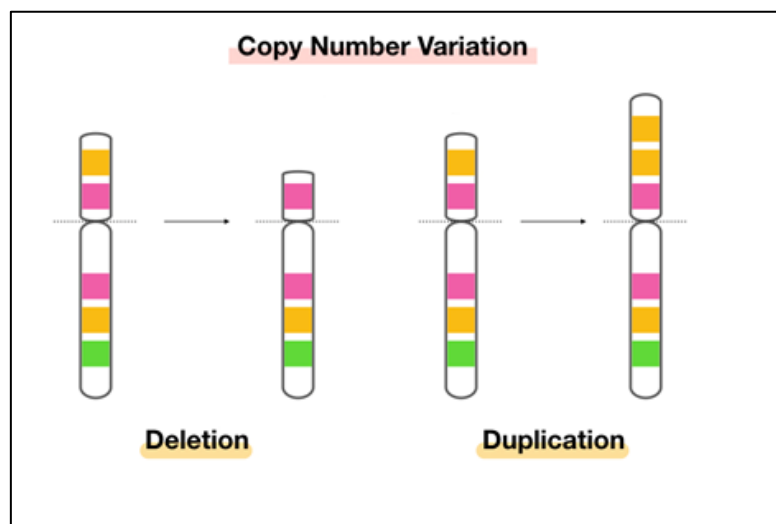


**Figura 1.** Descrição do Desequilíbrio de Ligação (Bush & Moore, 2012).

## 1.4. Copy Number Variation (CNV)

O GWAS têm sido extremamente bem-sucedidos em associar polimorfismos de nucleotídeo único (*Single nucleotide polymorphism* - *SNP*) com suscetibilidade a doenças comuns, porém as variantes genéticas do tipo SNP representam apenas uma fração do componente genético das doenças mais comuns. Outra classe de achados genômicos são as Variantes Estruturais (*Structural Variants* - *SV*). Por definição, as SVs são um grupo de variações genômicas que envolvem vários segmentos de DNA, com tamanho variando entre 50pb (pares de bases) até milhares de pb e podem envolver vários genes.

As SVs incluem vários tipos de variações, como: deleções, duplicações, inserções e translocações (Escaramís et al., 2015). Dentre as diversas classes de SVs, existem as variações no número de cópias (*Copy Number Variation* - *CNV*) que são ganhos ou perdas de segmentos de DNA (Weischenfeldt et al., 2013) e podem envolver até 12% do genoma humano (Redon et al., 2006). As variações de número de cópias podem ser relevantes depende da região, assim, uma caracterização detalhada destes eventos é de extrema importância (Weischenfeldt et al., 2013).



**Figura 2.** Exemplo de CNVs. Adaptado de <https://geneticeducation.co.in/what-is-copy-number-variation-and-how-to-detect-it>.

- **Mecanismos de Formação das CNVs**

A formação das CNVs pode ser resultado de eventos complexos como a recombinação, a replicação ou reparo do DNA. Quando ocorre algum problema nesses processos, isso pode originar alterações no número de cópias de alguns segmentos cromossômicos, de tal forma, que os



rearranjos podem se relacionar com alguma doença ou até mesmo dar origem a um polimorfismo populacional.

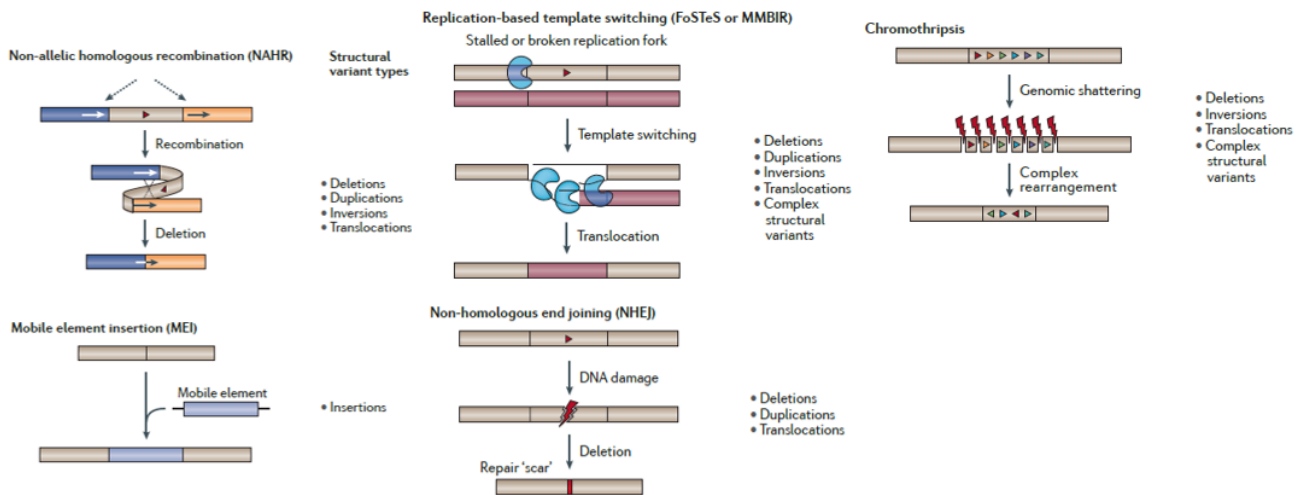
Dentre os diversos mecanismos moleculares envolvidos em recombinação, replicação e reparo do DNA, são descritos como eventos relacionados com a formação de CNVs: a *Non-allelic homologous recombination (NAHR)*, a *Fork Stalling and Template Switching (FoSTeS) / Microhomology-Mediated Break-Induced Replication (MMBIR)*, a *Nonhomologous DNA end joining (NHEJ)* e a *Chromothripsis* (Weischenfeldt et al., 2013) e (Lupski, 2016). A Figura 3 ilustra a formação desses eventos.

A *Non-allelic homologous recombination (NAHR)* envolve recombinação entre dois segmentos de DNA que possuem alta similaridade, conhecidos como *Low Copy Repetition (LCR)*. O NAHR tem como função restaurar quebras cromossômicas, contudo, quando ocorre uma falha resulta em alguns rearranjos genômicos. Estima-se que de 10% a 22% de todos os rearranjos do genoma em humanos, das SV de modo geral, sejam resultados deste evento (Parks et al., 2015). Os *Low Copy Repetition (LCR)* constituem até 5% do genoma humano haplóide de referência. Esses segmentos de DNA são regiões previamente duplicadas ao longo da evolução e, portanto, têm poucas repetições de cópia. Quando as regiões de LCRs estão localizados a uma distância inferior a 10 Mb pode ocorrer o desalinhamento de cromossomos ou cromátides e cruzamentos desiguais, originando deleções ou outras duplicações do DNA, promovido pelo NAHR (Lupski, 2016).

O *Fork Stalling and Template Switching (FoSTeS) / Microhomology-Mediated Break-Induced Replication (MMBIR)* é um evento mutacional em que ocorrem rearranjos a partir da replicação de segmentos cromossômicos. Nestes rearranjos a replicação é induzida por quebras mediadas por regiões com microhomologia (MMBIR) e frequentemente originam duplicações (J. A. Lee et al., 2007). A geração de CNVs por esse evento, pode ocorrer durante a meiose ou na mitose. Logo, mesmo gêmeos monozigóticos podem apresentar diferenças para essas variantes ou ocasionar, de modo geral, que indivíduos apresentem as variantes em mosaico entre os tecidos e mesmo dentro dos tecidos (Escaramís et al., 2015).

Muitas variantes do tipo CNV também podem ser atribuídas pelo mecanismo de *Nonhomologous End Joining (NHEJ)*. O NHEJ é um processo que repara quebras de fita dupla de DNA na ausência de regiões de extensa homologia. Ao detectar quebras de dupla fita no DNA, através do NHEJ as extremidades são conectadas, modificadas e religadas. Ao ocorrer um erro, durante esse processo pode ser adicionado ou retirados de alguns nucleotídeos e gerar uma “cicatriz de reparo”. Contudo, essa modificação equivocada provoca em evento mutacional (Weischenfeldt et al., 2013) (Lupski, 2016). A *Chromothripsis* é um evento mutacional que envolve a quebra cromossômica em vários pontos, seguindo pelo reparo de DNA. Na formação tumoral, a *Chromothripsis* mostra-se como uma fonte de rápida aquisição de mutações, contudo também está

relacionado a mutações em linhagem germinativa (Cortés-Ciriano et al., 2020) (Kloosterman et al., 2011).



**Figura 3.** Mecanismos de formação das SVs. Adaptado (Weischenfeldt et al., 2013)

• **Metodologias de Detecção de CNVs**

Existem várias técnicas para detectar CNVs (Tabela 1), como a Hibridação in situ por Fluorescência (*Fluorescence In Situ Hybridization - FISH*), o Microarranjo de DNA (*Microarray*) e abordagens baseadas em Sequenciamento de Nova Geração (*Next Generation Sequencing - NGS*) como painéis, Sequenciamento do Exoma Completo (*Whole Exome Sequencing - WES*) e Sequenciamento do Genoma Completo (*Whole Genome Sequencing - WGS*). Dentre elas, a técnica de *Microarray* é uma das mais robustas e utilizadas, em que a partir da chamada de genótipos de marcadores é possível inferir com confiança o número de cópias de segmentos cromossômicos ao longo do genoma (Carter, 2009).

**Tabela 1.** Breve descrição dos principais métodos de detecção de CNVs. Adaptado (Pös et al., 2021)

Método	Nível de Detecção	Técnica
Cariótipo e FISH	CNV grandes	Banda G e Hibridização in situ
MLPA	CNV pequenas	PCR multiplex
CGH array	CNV médias e grandes	Hibridização Genômica Comparativa
SNParray	CNV pequenas a grandes	Hibridização e detecção por SNPs
Sequenciamento Painéis	CNV pequenas a grandes	Sequenciamento por NGS
Sequenciamento Exoma (WES)	CNV pequenas a grandes	Sequenciamento por NGS

As plataformas de *Microarray*, como as de *SNParray* (microarranjo baseado em marcadores do tipo SNP), se tornaram ainda mais robustas, o que tem facilitado a detecção confiável de CNVs. Utilizando essa abordagem, as CNVs são captadas a partir de dados de marcadores (SNP) ao longo do genoma onde são calculadas métricas como a *Log R Ratio - LRR* (Razão Log R) da intensidade do sinal do marcador e o *B Allele Frequency – BAF* (Frequência do Alelo B) (Weischenfeldt et al., 2013). Calcular essas métricas frente aos avanços das plataformas de *Microarray* envolve o uso de diferentes ferramentas de bioinformática e propor etapas de processamento para a identificação acurada de CNVs ainda é um desafio no processamento de dados genéticos.

- **CNVs em estudos de associação**

As CNVs podem ter origem herdada ou *de novo*, e dependendo da região onde ocorrem, podem interferir na composição genética e em funções biológicas. Além disso, a presença de CNVs pode resultar em amplificação ou regulação negativa de genes sensíveis à dosagem, o que pode contribuir para variações na suscetibilidade a doenças e caracterizar fatores de riscos (K. W. Lee et al., 2012).

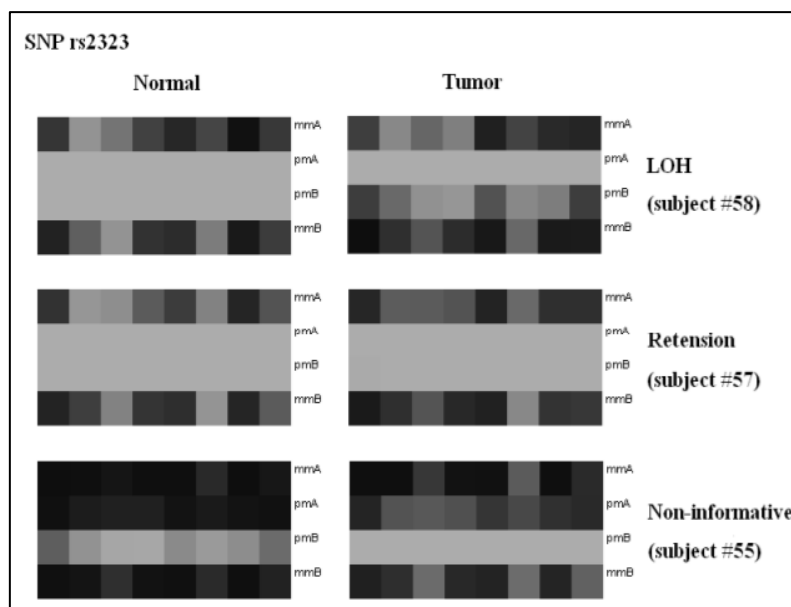
Com base nisso, vários trabalhos associam as CNVs a diversos tipos de fenótipos desde de doenças monogênicas, bem como multifatoriais (como as doenças crônicas não-transmissíveis), correlacionando as CNVs na inferência de fatores de risco (Almal & Padh, 2012). Além de conhecer as regiões com variação no número de cópias para elucidação de doenças humanas, a detecção de CNVs também é importante para estudar a diversidade genômica entre populações (Lin et al., 2013).

Nesse contexto, a inclusão das informações de CNVs nos estudos de GWAS, tem expandido a possibilidade de identificação de variantes moleculares com influência em fenótipos. Isso tem contribuído para desvendar a “herdabilidade perdida”, supostamente presente em muitos estudos de mapeamento por falta de análises enriquecidas de informação sobre diferentes tipos de variações moleculares (K. W. Lee et al., 2012).

## 1.5. **Loss of Heterozigosity (LOH)**

*Loss of heterozygosity (LOH)* é uma instabilidade cromossômica, caracterizada pelo desequilíbrio alélico, onde uma região genômica heterozigota se torna homozigota através da perda de um dos alelos (*Copy Loss - LOH*) e possível duplicação do mesmo segmento (*Copy Neutral - LOH*). As regiões homozigotas podem representar subpopulações, determinar nível de

consanguinidade e diversos estudos correlacionam as regiões LOH como uma das principais causas de tumorigênese ((Ryland et al., 2015) e (Zhang & Sjöblom, 2021)). A Figura 4 ilustra um evento de LOH.



**Figura 4.** Exemplo de trechos de LOH em um marcador em células normais e tumorais. Adaptado (Lin et al., 2004)

No contexto oncológico, os fatores que conferem a diversidade e heterogeneidade dos tumores malignos, principalmente o câncer de mama, são as variações genéticas, decorrentes do polimorfismo gênico e, em especial, ao LOH. Foi demonstrado que o LOH em alguns genes pode ser um bom marcador prognóstico (Deryusheva et al., 2017).

Além da correlação com o câncer, outra aplicação para análise de LOH é a observação de mutações do tipo perda de função (*Loss of Function* - LOF). (Narasimhan et al., 2016) observou pacientes com alta taxa de consanguinidade e verificou alelos homozigotos em regiões que deveriam conferir doença genética, e até mesmo uma família com a ausência do gene PRDM9 por recombinação, fenômeno conhecido como *knockout* gênico. Entender as correlações de *knockout* gênico, ajudam a compreender a função dos genes, suas interações e os mecanismos das doenças em humanos.

Mesmo com diversas aplicações em doenças complexas, a investigação de LOH ainda está em grande parte inexplorada, contudo, estima-se que seu impacto potencial em doenças complexas é enorme (Lencz et al., 2007). Logo, é de extrema importância a geração de bancos de dados abrangendo estas variações no genoma humano, criando a oportunidade de desvendar novos sítios de doenças.

## 2. Objetivos

O objetivo primário do presente estudo é propor um processamento customizado para identificar regiões de variação no número de cópias (CNV) e regiões com perda de heterozigosidade (LOH) a partir de dados de *SNParray*. Com base nos dados do projeto ISA-Nutrição (Fisberg et al., 2018) será construído um banco inicial de referência de CNVs e LOHs para a população brasileira. Considerando a heterogeneidade de nossa população devido à sua recente miscigenação, um primeiro passo será caracterizar a ancestralidade global dos participantes. Com as regiões de CNVs e LOHs identificadas, como objetivo secundário, alguns polimorfismos alvo associados à susceptibilidade a doenças crônicas localizados nessas regiões, serão estudados e associados a fenótipos de interesse.

## 3. Materiais e Métodos

### 3.1. Banco de Dados

Os dados utilizados neste trabalho são de uma subamostra do projeto Inquérito de Saúde de São Paulo (ISA-Capital), coordenado pelo Professor Chester Luiz Galvão César, da Faculdade de Saúde Pública da Universidade de São Paulo, em parceria com a Secretaria Municipal de Saúde de São Paulo. O ISA-Capital tem como objetivo avaliar as condições de vida, situação de saúde e uso de serviços de saúde em uma amostra de moradores do município de São Paulo. Em suas diferentes versões, o projeto conta com apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processos 98/14099-7; 2007/51488-2; 2009/15831-0; 2012/22113-9) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, processos 502948/2003-5, 481176/2008-0; 472873/2012-1) e aprovação pelos Comitês de Ética em pesquisa da Secretaria Municipal de Saúde e da Faculdade de Saúde Pública (Alves et al., 2018).

Os participantes do ISA-Capital são avaliados periodicamente com dados desde o ano de 2003. No ano de 2015, alguns participantes do projeto, além de participarem da linha de base do inquérito (n = 4.043), participaram de outras duas etapas subsequentes: aplicação do recordatório alimentar (n = 1.737) e coleta de amostra sanguínea, aferição da pressão arterial e avaliação antropométrica (n = 901), Tabela 2. A amostragem dos indivíduos foi realizada por conglomerados e estratificada em dois estágios (setores censitários urbanos e domicílios), e todos os residentes com 12 anos ou mais foram convidados a participar do estudo, o que derivou o projeto ISA-Nutrição que é coordenada pela Professora Regina Mara Fisberg.

O objetivo do ISA-Nutrição é avaliar o consumo alimentar e outros fatores de estilo de vida modificáveis em uma subamostra do ISA-Capital, com indivíduos com 12 anos ou mais e, no caso de mulheres, que não estivessem grávidas ou amamentando no período da coleta. Em 2015, o ISA-Nutrição contou com apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo 2012/22113-9 e projeto temático 2017/05125-7) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, processos 472873/2012-1; 402674/2016-2; 303904/2021-6) e aprovação pelos Comitês de Ética em pesquisa da Secretaria Municipal de Saúde e da Faculdade de Saúde Pública (COEP nº 275/09; CAAE nº 30848914.7.0000.5421). (Fisberg et al., 2018).

**Tabela 2.** Descrição dos participantes do projeto *ISA-Nutrição* por gênero e faixa etária.

<b>Participantes ISA-Nutrição</b>			
<b>Faixa etária</b>	<b>Masculino</b>	<b>Feminino</b>	<b>Total</b>
<b>12 a 20 anos</b>	141	150	291
<b>20 a 60 anos</b>	153	149	302
<b>60 ou mais</b>	152	156	308
<b>Total</b>	446	455	901

No ISA-Capital 2015, foram realizadas 865 entrevistas com adolescentes (12 a 19 anos), 2.169 com adultos (20 a 59 anos) e 1.025 com idosos ( $\geq 60$  anos) ( $n = 4.059$ ). A idade de início dos adolescentes foi definida de acordo com o Estatuto da Criança e do Adolescente (BRASIL, 1991), que considera os adolescentes a partir de 12 anos, e as demais seguiram a Organização Mundial da Saúde (de Onis et al., 2007), e o Estatuto do Idoso (BRASIL, 2003).

### **3.2 Categorização dos participantes**

- **Declaração de Raça**

No Brasil, o Instituto Brasileiro de Geografia e Estatística (IBGE) utiliza os termos de raça-autodeclarada como classificação dos indivíduos em pesquisas, em que os próprios indivíduos procuram uma classificação à qual melhor se definem. São utilizadas cinco categorizações arbitrárias: branca, parda, preta, amarela e indígena. No questionário do projeto ISA-Capital além das cinco categorias fechadas, os indivíduos poderiam definir outra. Neste caso, foram utilizados termos como: “moreno”, “moreno claro”, “moreninho” e não declarado. As derivações são comuns em populações miscigenadas como a brasileira, devido à diversidade populacional e o caráter observacional da categorização. Para o estudo, as derivações da classificação “moreno” foram acrescentadas à categoria parda

- **Local de Nascimento**

No questionário também foi abordada a naturalidade dos participantes, com o objetivo de observar a grande diversidade de origem da atual população paulistana. O local de nascimento foi distribuído em quatro grandes grupos: (1) nascidos na cidade de São Paulo, (2) nascidos em outra cidade do Estado de São Paulo, (3) nascidos em outro Estado do Brasil e (4) estrangeiros. Quando aplicável, procurou-se a especificação de qual a cidade ou região de nascimento dos indivíduos. A

Tabela 3 descreve os dados relativamente à faixa etária, gênero, raça-autodeclarada e local de nascimento.

**Tabela 3.** Informações sobre os participantes do Projeto ISA-Nutrição

	ISA-Nutrição	
	n	%
<b>População Total</b>	841	100
<b>Faixas Etárias</b>		
Adolescente (12  -- 20 anos)	250	29,7
Adulto (20  -- 60 anos)	289	34,4
Idoso (60  -- 94 anos)	302	35,9
<b>Gênero</b>		
Feminino	423	50,3
Masculino	418	49,7
<b>Raça auto-declarada</b>		
Branca	422	50,2
Preta	84	10
Parda	309	36,7
Amarela	13	1,6
Indígena	2	0,2
Não Informada	11	1,3
<b>Local de Nascimento</b>		
Cidade de São Paulo (Região Sudeste)	435	51,7
Estado de São Paulo (Região Sudeste)	94	11,2
Outro Estado da Região Sudeste	55	6,5
Estados da Região Norte ou Centro-Oeste	8	1
Estado da Região Nordeste	187	22,2
Estado da Região Sul	24	2,9
Outro Estado do Brasil (não especificado)	5	0,6
Outro País	26	3,1
Não Informado	7	0,8

### 3.2.1. Extração e métricas de qualidade do material genético

Para a obtenção do material genético de cada participante do ISA-Capital 2015, foram realizadas coletas de sangue no domicílio de cada indivíduo, por um técnico de enfermagem, instruído a seguir os procedimentos padronizados de coleta de amostras biológicas do estudo. Os indivíduos foram instruídos sobre os preparos de coleta e então foram coletados aproximadamente 40 mL de sangue



venoso em tubos com EDTA (ácido etileno-diamino-tetracético) e seco para utilização de testes bioquímicos.

O material biológico foi submetido à extração de DNA através do protocolo automatizado pelo equipamento *QIAasymphony SP Biorobot* com o kit de extração *QIAasymphony DNA MID Kit 96* (*Qiagen, Hilden, Alemanha*) no Centro de Pesquisa de Genoma Humano e Células-Tronco do Instituto de Biociências da Universidade de São Paulo (IB-USP). Dentre as 901 amostras, 40 apresentaram menos de 1 mL de sangue periférico, o que tornou a extração automatizada inviável, logo para estas amostras, adotou-se o protocolo de extração manual. Contudo, 1 amostra ainda não atingiu os parâmetros mínimos para o protocolo e foi excluída na análise.

Para a checagem da qualidade do material genético extraído, as amostras foram avaliadas através do equipamento *NanoDrop™ 2000* (*Thermo Fisher Scientific, Waltham, EUA*), que, por meio do método de espectrofotometria avalia o índice de DNA e proteínas no resultado da extração, inferindo a pureza e a qualidade do método. Após essa avaliação, 130 amostras não atingiram as métricas adequadas e foram submetidas ao protocolo de purificação, com o objetivo de reduzir a taxa de proteínas pela desnaturação. Ainda, para confirmar a qualidade e integridade do material, foi realizada uma quantificação por fluorescência, selecionando aleatoriamente amostras extraídas pelo método automatizado e manual. O procedimento foi realizado em uma corrida de eletroforese, onde por 40 minutos em um gel de agarose de 1% a 100V, marcado com *SYBR™* (*Thermo Fisher Scientific*), as amostras foram testadas para mostrar o padrão de faixas e peso molecular. Ao final 900 amostras se apresentaram adequadas para a genotipagem.

### **3.2.2. Plataforma de Genotipagem: *Axiom™ 2.0 Precision Medicine Research Array***

Para a genotipagem das amostras foi utilizada a plataforma *Axiom™ 2.0 Precision Medicine Research Array* (*Thermo Fisher Scientific*). A plataforma utiliza 902.527 marcadores *SNP* (*Single Nucleotide Polymorphism*) que podem ser aplicados a diversos tipos de estudo como (Figura 5): (1) estudo de associação do genoma (GWAS), (2) variantes clinicamente acionáveis, (3) Variantes do fenótipo sanguíneo, (4) Variantes farmacogenômicas, (5) Marcadores comuns do câncer, (6) Marcadores relacionados à imunidade, (7) Variantes funcionais e (8) Variantes de rastreamento de impressões digitais/amostras.

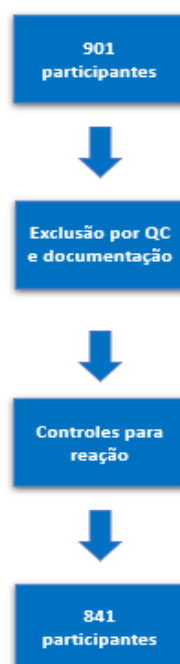
Para o GWAS, o kit utiliza cerca de 800.000 SNPs para contemplar as 5 superpopulações do Projeto 1000 genomas (*The 1000 Genomes Project Consortium, 2015*): Africana (AFR), Americana (AMR), Leste-asiática (EAS), Europeia (EUR) e Sul-asiática (SAS).

Variant category	Number of markers*
Genome-wide imputation grid	>800,000
NHGRI-EBI GWAS catalog	>15,000
ClinVar	>23,000
American College of Medical Genetics (ACMG, subset of ClinVar)	>9,000
Additional high-value markers (subset of ClinVar: <i>APOE</i> , <i>BRCA1/2</i> , <i>DMD</i> , <i>CFTR</i> )	>2,000
HLA	>9,000
KIR	>1,400
Autoimmune/inflammatory	>250
Pharmacogenomic	>1,200
Blood phenotype	>2,000
Common cancer variants	>300
Loss of function	>33,000
Expression quantitative trait loci (eQTL)	>16,000
Fingerprinting and sample tracking	>300
<b>Total markers</b>	<b>902,527</b>

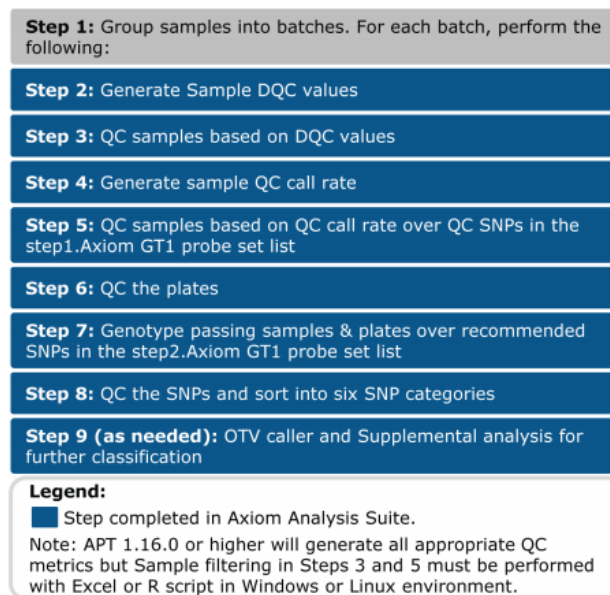
\* Content in categories may overlap.

**Figura 5.** Tabela com o grupo de marcadores da plataforma *Axiom PMRA*.

As amostras foram enviadas para o laboratório da *Affymetrix (Thermo Fisher Scientific)* em 9 placas com capacidade de 96 poços. Seguindo a recomendação da empresa, houve uma redução para 846 amostras no estudo, para alocar 18 poços da placa para controles. Para contemplar as 846 amostras, foram excluídos da análise os indivíduos com dados antropométricos ausentes, gestantes e prováveis irmãos. A Figura 6 apresenta os passos envolvidos na redução amostral do estudo. Após o processamento técnico e leitura dos *chips*, as amostras foram analisadas no programa *Axiom Analysis Suite* utilizando o protocolo *Best Practices Genotyping Analysis Workflow* (Figura 7).



**Figura 6.** Protocolo da seleção de amostra para genotipagem.




**Figura 7.** Etapas para as melhores práticas de genotipagem de análise de fluxo de trabalho.

Antes de iniciar o processamento das amostras na plataforma, uma das métricas utilizadas para inferir a qualidade é a DQC (*Dish Quality Control*). Como o ensaio de genotipagem é baseado em fluorescência, a DQC é a razão do contraste entre o canal das bases AT e o canal das bases GC em locais genômicos não polimórficos. Quando a qualidade da genotipagem está adequada, o sinal é alto no canal esperado pela sequência, contudo, quando o sinal é baixo, infere-se que a qualidade da amostra está igualmente baixa. A escala da DQC está entre 0 e 1, valores mais próximos do 0 indicam baixa resolução ou baixa qualidade e os valores mais perto de 1 indicam uma ótima resolução e, em consequência, uma amostra de boa qualidade. Seguindo o protocolo *Best Practices Genotyping Analysis Workflow*, recomenda-se a utilização de amostras com o DQC acima de 0,82. Todas as amostras atingiram o valor mínimo e iniciaram o processamento de análise. O cálculo da DQC é dado por:

$$DQC = \frac{AT_{signal} - CG_{signal}}{AT_{signal} + CG_{signal}}$$

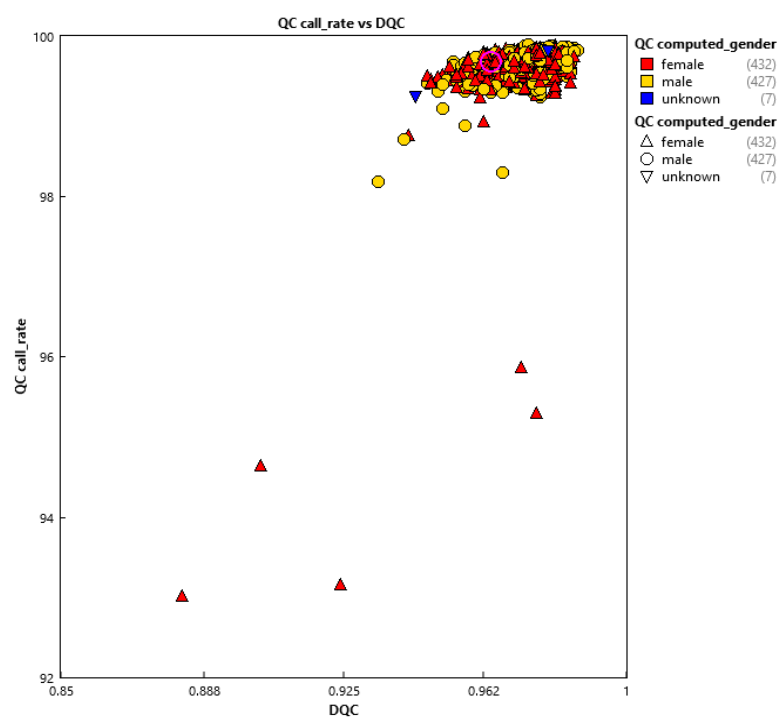
Para um bom desempenho do processo de análise, mais alguns parâmetros de qualidade devem ser seguidos para classificar as amostras. Mesmo com os valores de DQC adequados, algumas amostras podem ainda ter uma baixa qualidade para genotipagem. Para não gerar vieses no agrupamento das amostras, após a classificação DQC é calculada a *QC call rate*. Essa métrica consiste na genotipagem de cerca de 20.000 SNPs autossômicos previamente testados para verificar a performance das chamadas. Recomenda-se no protocolo *Best Practices Genotyping Analysis* o

valor mínimo de 97% de *QC call rate*. Das amostras estudadas, cinco foram excluídas por apresentarem valores inferiores ao ponto de corte (Figura 8).

Sample Filename	Pass/Fail 	DQC	QC call_rate
a550778-4400703-091721-547_D07.CEL	Fail	0.972	95.87
a550778-4400703-091721-547_F02.CEL	Fail	0.976	95.29
a550778-4405421-101021-811_E07.CEL	Fail	0.924	93.15
a550778-4405421-101021-817_F06.CEL	Fail	0.903	94.635
a550778-4405421-101021-817_H02.CEL	Fail	0.882	93.015

**Figura 8.** Amostras com *QC call rate* abaixo do limite de qualidade (97%)

Na Figura 9, ao comparar ambas as métricas através da visualização gráfica, é possível verificar as cinco amostras discrepantes em relação ao parâmetro *QC call rate* e o comportamento para a métrica DQC. Apesar de algumas amostras receberem valores bem altos de DQC, não apresentaram uma boa performance na genotipagem.



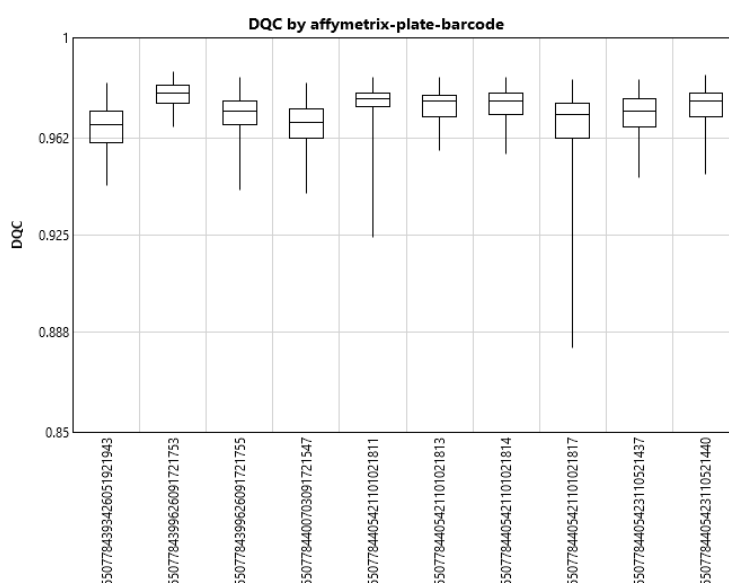
**Figura 9.** Gráfico *QC call rate* por DQC.

Outra etapa do protocolo *Best Practices Genotyping Analysis* é a avaliação da placa, ou seja, a avaliação do efeito de lote de processamento das amostras. É importante ressaltar que a primeira

recomendação para o controle do efeito de placa é a aleatorização das amostras às placas e às correspondentes posições dentro das placas. A aleatorização evita a presença de fontes de variação indesejáveis e de confundimento. A avaliação do efeito de placa é importante, tendo em vista que intensidades diferentes entre placas podem gerar erros entre alguns conjuntos de sondas, o que pode contribuir para uma clusterização (agrupamento) errada. Os impactos são muito relevantes, como provocar desvios na Probabilidade do Menor Alelo (*Minor Allele Frequency – MAF*) em alguns SNPs pela comparação entre outras placas com melhor performance. Isso aumenta as taxas de falso-positivo no experimento, além de diminuir a acurácia da genotipagem. A métrica *Plate Pass Rate* é a média entre as amostras que passaram no controle de qualidade individual pelo total de amostras da placa. O valor mínimo adequado é 98,5% de *Plate pass rate* sendo seu cálculo dado por:

$$Plate\ pass\ rate = \frac{Samples\ passing\ DQC\ and\ 97\%\ QC\ call\ rate}{Total\ sample\ on\ the\ plate} * 100$$

Durante a execução do protocolo, nenhuma placa foi excluída, e a maior variabilidade foi restrita apenas às placas onde tiveram algumas amostras que não atingiram a métrica individual de *QC call rate*. As Figuras 10 e 11 apresentam os diagramas de caixa (*boxplots*) com as distribuições observadas da métrica DQC e da *QC call\_rate*, respectivamente, nas amostras para as 9 placas do estudo.



**Figura 10.** *Boxplot* (Diagrama de Caixa) da métrica DQC para as 9 placas processadas.

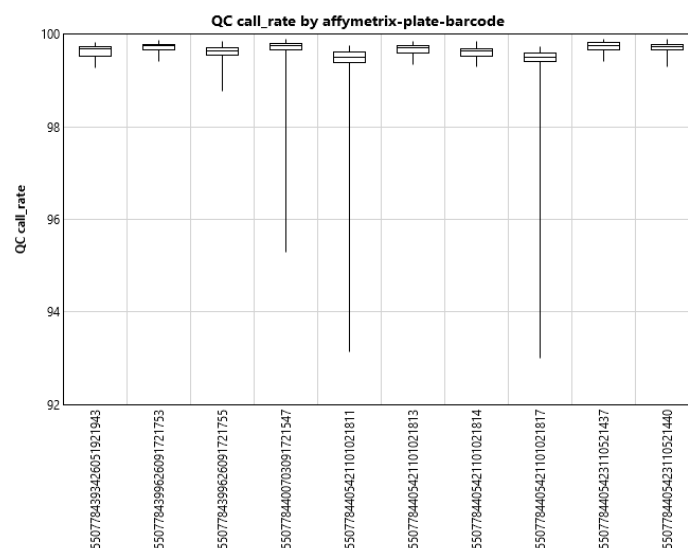


Figura 11. *Boxplot* (Diagrama de Caixa) da métrica *QC call rate* para as 9 placas processadas.

### 3.3. Genotipagem

- **Algoritmo *AxiomGT1***

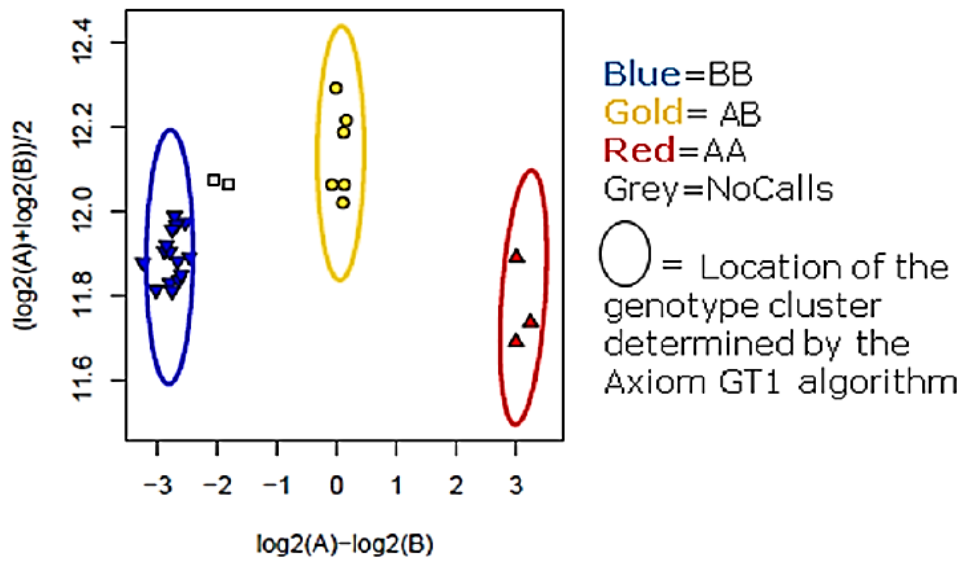
No protocolo de genotipagem foi utilizado o algoritmo de cluster (agrupamento) de genotipagem *AxiomGT1*, baseado no algoritmo BRLMM – P (Modelo Bayesiano Robusto Linear com Genótipo Classificador de Distância de Mahalanobis). Esta versão do algoritmo atualiza as localizações *a priori* de agrupamentos de genótipos a partir dos dados do estudo, e calcula três localizações de agrupamentos *a posteriori*. Para cada SNP, é construído o gráfico de dispersão bidimensional das intensidades de expressão (log-transformadas) das bases (ou alelos) A e B, no qual, as localizações dos agrupamentos de genótipos são definidas por elipses de concentração correspondentes às distribuições de genótipos AA, AB e BB. *As prioris* podem ser genéricas, o que significa que o mesmo local preposicionado é fornecido para cada SNP, ou SNP específico.

Basicamente, o agrupamento *AxiomGT1* é realizado em duas dimensões com a log-transformação e rotação das intensidades de expressão dos alelos A e B (Figura 12). A dimensão Y é calculada como:

$$Y(size) = [\log_2(A) + \log_2(B)]/2$$

e a dimensão X é calculada como:

$$X(\text{contrast}) = \log_2(A) - \log_2(B)$$



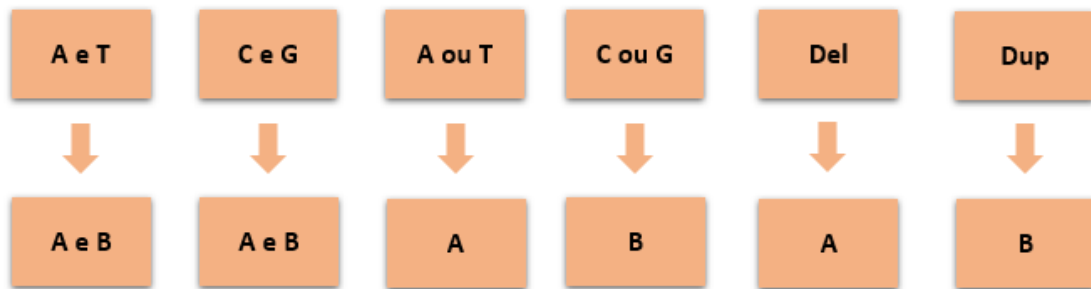
**Figura 12.** Log-Transformação e Rotação de A e B, aplicados pelo algoritmo AxiomGT1.

A dimensão X carrega as principais informações para distinguir agrupamentos de genótipos. A dimensão X é chamada de contraste ou desvio (*contrast*) e a dimensão Y é chamada de tamanho ou média (*size*) das leituras das bases A e B. Os gráficos de agrupamento são produzidos pelo algoritmo *SNPlisher*.

Os genótipos do tipo *NoCalls* (sem chamada) são produzidos pelo algoritmo AxiomGT1 em amostras cujas pontuações de confiança estão acima do limite de padrão de 0.15. A pontuação de confiança é essencialmente a diferença (1 - probabilidade posterior do ponto pertencer ao grupo de genótipos atribuído). Assim, as pontuações de confiança variam entre zero e um, e pontuações de confiança mais baixas indicam chamadas de genótipos mais confiantes, acima desse valor, a chamada de genótipo para a amostra é convertida para o tipo *NoCall*.

Na genotipagem a referência de cada marcador SNP obedece à nomenclatura “*affy\_snp\_id*” (exemplo: Affx-19965213) e os trechos do DNA cujas intensidades combinadas identificam o marcador como “*probsets*” (exemplo: probset\_id = AX-33782819). Nos bancos de dados é possível encontrar também a correspondência para o banco de dados do *dbSNP* (Banco de Dados de Polimorfismo de Nucleotídeo Único), com a identificação “*rs\_id*”. Nem todos os marcadores do kit têm correspondência com os números de identificação do banco *dbSNP*, tendo em vista que durante o desenvolvimento da tecnologia alguns SNPs foram colocados de forma customizada pela empresa *Thermo Fisher*.

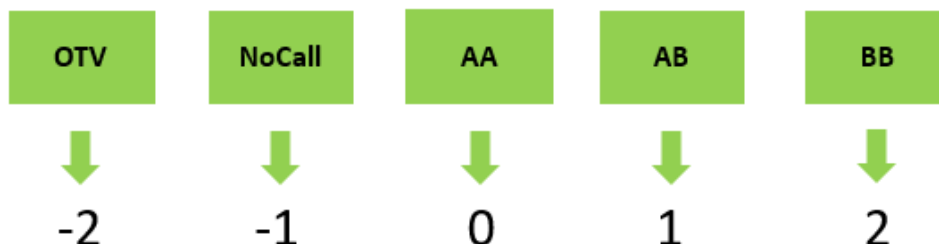
Para a definição dos alelos A e B, quando são A/T ou C/G, o critério definido é a ordem alfabética. Logo, a base Adenina é nomeada A e a Timina é nomeada B e/ou a Citosina é nomeada A e Guanina é nomeada B. Contudo, para os SNPs que não são A/T e C/G, o alelo A será Adenina ou Timina e o alelo B será Citosina ou Guanina. Para representar marcadores do tipo indel (deleções ou inserções) a deleção é definida como A e a base de inserção na fita anterior (*forward*) é definida como B. A Figura 13 ilustra essa conversão.



**Figura 13.** Conversão dos alelos para a nomenclatura A e B.

A codificação adotada pelo protocolo de genotipagem em cada marcador é (-2) para os marcadores em regiões fora do alvo (*off-target* - OTV), (-1) para os marcadores *NoCall*, (0) para os marcadores em homozigose AA, (1) para os marcadores em heterozigose AB e (2) para os marcadores em homozigose BB (Figura 14). Os OTV são marcadores com sequências significativamente diferentes das genotipadas pelo *Microarray*, formando um quarto cluster de indivíduos, diferentes dos classificados como AA, AB ou BB. Os marcadores *NoCall* são valores de intensidade inespecíficos ou marcadores que não passaram no controle de qualidade e um exemplo, é a detecção de chamadas na região do cromossomo Y quando o indivíduo é do sexo feminino.

Além das interpretações dos genótipos pelas classes acima, podem ocorrer representações diferentes, relacionadas a SNPs multialélicos, ou seja, que possuem mais de uma possibilidade de alelo alternativo ou variação no número de cópias do segmento (CNV), que pode ser uma duplicação ou deleção.



**Figura 14.** Valor de cada genótipo para o algoritmo de genotipagem.



### 3.4. Verificação de Qualidade dos SNPs

Após a genotipagem ocorre uma verificação de qualidade de cada marcador considerando as chamadas no nível populacional. Nesta etapa são calculadas 17 métricas básicas de qualidade, dentre elas: Frequência do Menor Alelo (*Minor Allele Frequency - MAF*), o valor p dos testes de genotipagem, o valor da estatística qui-quadrado para o equilíbrio de *Hardy-Weinberg* e p valor *para Hardy-Weinberg*. A partir das métricas básicas de qualidade os marcadores são classificados em 6 grupos (ver ilustração nas Figuras 15, 16 e 17).

- **PolyHighResolution:** marcador com boa resolução na clusterização e pelo menos dois exemplos do alelo menor frequência.
- **MonoHighResolution:** marcador com uma possível fusão na clusterização. Apresenta menos de dois exemplos do alelo menor frequência, que geralmente ocorre devida as amostras ter valores baixos de MAF.
- **NoMinorHom:** formação de dois clusters sem exemplos de genótipos homocigotos do alelo de menor frequência.
- **CallRateBelow:** a taxa de chamada do marcador está abaixo do limite, mas outras propriedades de cluster estão acima do limite.
- **Off-Target Variant:** marcadores em regiões fora do target, podem ser avaliados pela genotipagem OTV, contudo, não é indicado pelo workflow *Best Practices Genotyping Analysis*.
- **Other:** marcadores com uma ou mais propriedades de cluster que estão abaixo do limiar, ou seja, genótipos de menor qualidade.

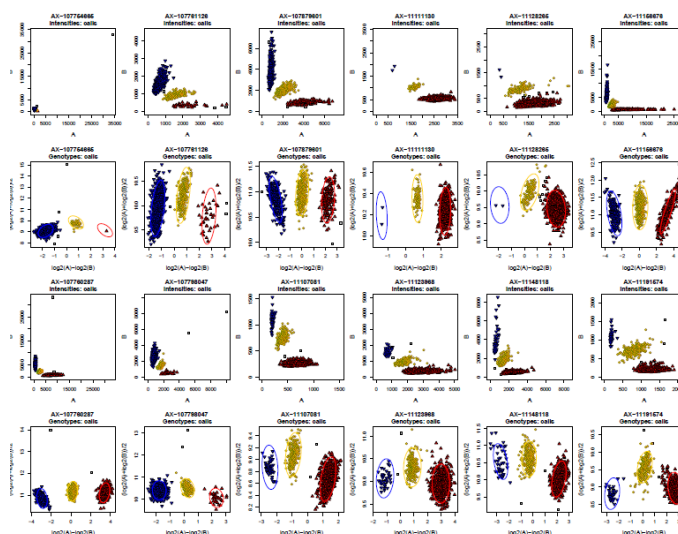
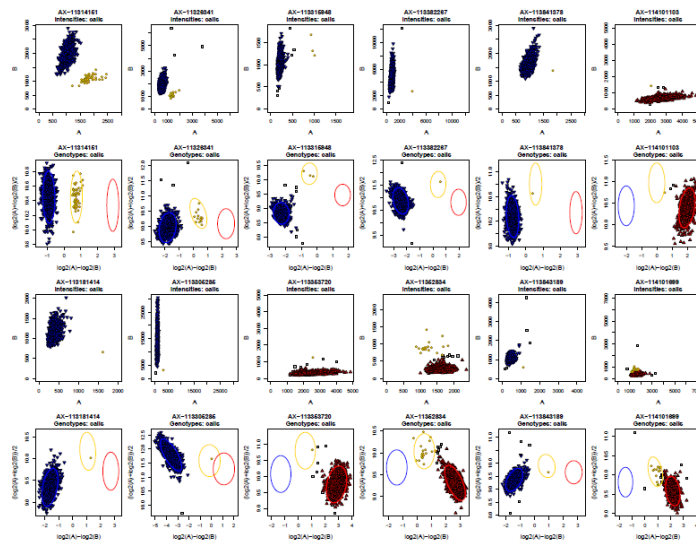
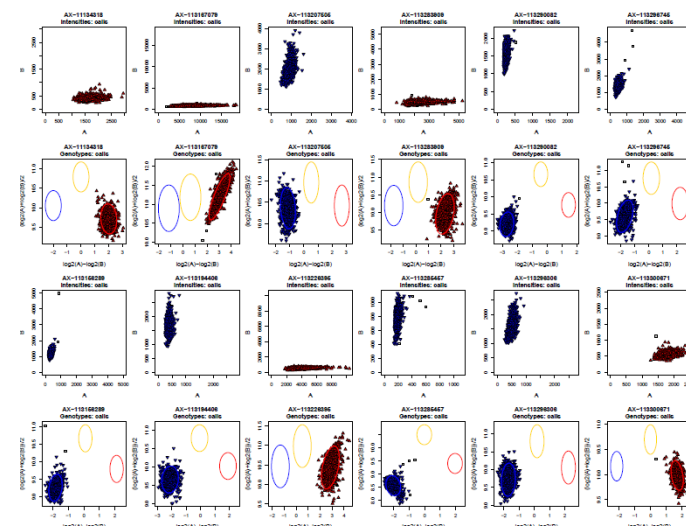


Figura 15. Exemplos de marcadores *PolyHighResolution*. Fonte: *User Guide: Axiom™ Genotyping Solution Data Analysis (Thermo Fisher Scientific)*



**Figura 16.** Exemplo de marcadores *NoMinorHom*. Fonte: *User Guide: Axiom™ Genotyping Solution Data Analysis (Thermo Fisher Scientific)*



**Figura 17.** Exemplos de marcadores *MonoHighResolution*. Fonte: *User Guide: Axiom™ Genotyping Solution Data Analysis (Thermo Fisher Scientific)*

Ao final da execução do protocolo *Best Practices Genotyping Analysis*, os marcadores das classes *PolyHighResolution*, *MonoHighResolution* e *NoMinorHom*, bem como os marcadores em hemizogose, são descritos em uma lista de SNPs recomendados para as análises, “Recommended.ps”, que, para os dados do ISA-Nutrição 2015, totaliza o número de 873.177 marcadores (Tabela 4).

**Tabela 4.** ISA-Nutrição - Distribuição dos marcadores por classe de qualidade.

<b>Classificação</b>	<b>Número de SNPs</b>
<i>PolyHighResolution</i>	448.809
<i>NoMinorHom</i>	324.531
<i>MonoHighResolution</i>	112.294
<i>CallRateBelowThreshold</i>	1.500
<i>Other</i>	12.428
<i>OTV</i>	2.998

### 3.5. Gráfico *Manhattan*

Para uma visualização generalizada dos genótipos, utilizamos o pacote *qqman* da linguagem R (D. Turner, 2018) para produzir um gráfico do tipo *Manhattan*, amplamente utilizado em estudos genômicos para mostrar a dispersão dos marcadores ao longo dos cromossomos de acordo com o valor p obtido em um teste estatístico. No gráfico *Manhattan*, por exemplo, os valores p representam o nível descritivo do teste de Equilíbrio de *Hardy-Weinberg*, extraídos dos arquivos de saída do protocolo *Best Practices Genotyping Analysis* e identificados pelo identificador *rs\_id*.

### 3.6. Filtragem dos SNPs

#### Filtragem pelo *Apt-Genotype-Axiom* e conversão para formato **PLINK**

A matriz de genótipos está representada no arquivo “AxiomGT1.call.txt”. Para o refinamento das análises, no caso dos dados ISA-Nutrição, foi necessária a aplicação de alguns filtros de controle de qualidade a partir das informações obtidas durante a execução do protocolo de genotipagem e outros filtros customizados, para a criação de um subconjunto mais adequado para a análise de ancestralidade global e para a criação de um banco de dados compartilhado para o projeto. A criação desse fluxo de critérios é de extrema importância para minimizar potenciais falsas descobertas (resultados falso-positivos).

Na execução de cada etapa, customizamos protocolo de processamento na linguagem *shell*, baseado no algoritmo *apt-genotype-axiom* versão 2.11.3 (*Thermo Fisher Scientific*), disponibilizado para a utilização do kit, e do algoritmo *PLINK* nas versões 1.9 e 2.0 (Chang et al., 2015). O *PLINK* é uma ferramenta gratuita e está disponível sob a licença GPLv3, permitindo o gerenciamento e a análise de dados do tipo SNP baseados pela posição genômica e pela descrição dos alelos.

A matriz de genótipos foi filtrada pela lista “Recommended.ps” com 873.177 marcadores de SNPs recomendados no protocolo *Best Practices Genotyping Analysis*, pela seleção das 841 amostras que atingiram as métricas de qualidade e pela seleção de todos os SNPs com identificação para o banco de dados do *dbSNP* com seus respectivos identificadores *rs\_id*. Ao final restaram 838.143 marcadores e os dados foram convertidos em arquivos (\*.ped) com a informação dos genótipos dos 841 indivíduos e (\*.map) com as informações sobre os 838.143 marcadores, para compatibilidade com o software *PLINK*.

## **Filtro por Regiões Complexas**

Removemos SNPs de regiões genômicas complexas, como as regiões palindrômicas e dos genes HLA. As regiões palindrômicas estão presentes no genoma de vários organismos e podem estimular deleções durante a replicação do DNA e a recombinação intercromossômica entre sequências homólogas, sendo responsáveis por 83% das deleções e pequenas inserções no genoma (Ganapathiraju et al., 2020). SNPs nessas regiões possuem alta variabilidade de frequências alélicas. Ainda, excluímos os SNPs da região HLA (*Human leukocyte antigen* - antígeno leucocitário humano) localizada no cromossomo 6p21.3. A região HLA é considerada uma das regiões mais polimórficas do genoma humano. Os genes HLA têm vários alelos que apresentam ampla variação nas populações humanas (Pillai et al., 2014). A exclusão desses marcadores, especificamente para objetivo de caracterizar a ancestralidade da amostra, foi realizada pelo protocolo de processamento utilizando o algoritmo *PLINK 1.9*.

## **Filtro por Frequência**

- ***Minor Allele Frequency (MAF)***

A Probabilidade do Alelo Menor (*Minor Allele Frequency* - MAF) é um importante parâmetro para estudos genéticos, fornecendo informações relevantes na diferenciação de variantes comuns e raras na população estudada. Os marcadores com valor de MAF muito baixo estão mais propensos a erros, que podem ocorrer pelo menor número de amostras representando os agrupamentos de genótipos e limitando a análise pela baixa acurácia da maioria dos algoritmos em alelos raros (Pongpanich et al., 2010). Neste estudo utilizamos um limite para os valores de MAF < 0,0001, um valor de flexível para a construção de um banco de dados primário, filtrando através do valor de frequência na ferramenta *PLINK*.

- ***Equilíbrio de Hardy-Weinberg (HW)***

O teste para equilíbrio de *Hardy-Weinberg* (HWE) é uma medida que demonstra a diferença entre as proporções observadas de chamadas heterozigotas em uma população, e a razão das proporções

esperadas sob independência alélica. O teste deve ser realizado em indivíduos não relacionados e com ascendências relativamente homogêneas. Os marcadores SNPs com desvios extremos de HWE normalmente são usados para filtrar o conjunto de dados, pela capacidade do teste em demonstrar erros grosseiros de genotipagem, quando o valor  $p$  do teste de HWE é menor do que um limite apropriado para testes múltiplos, representados em valores extremos. Os valores  $p$  foram calculados nos marcadores SNPs identificando pela classificação “*probeset*”, e definimos o valor de filtragem com um valor  $p$  do HWE  $< 0,0001$ , um valor que também torna o banco primário mais flexível, com a ferramenta PLINK. Para a avaliação da independência alélica dos marcadores, ou seja, se estão sob o equilíbrio de *Hardy-Weinberg*, utilizamos o teste Qui-Quadrado (indicado na Seção 1.3)

- **Desequilíbrio de Ligação - LD**

O Desequilíbrio de Ligação (LD) é explorado para otimizar estudos genéticos, para evitar a genotipagem de SNPs que fornecem informações redundantes, tendo em vista, que é realizada uma avaliação para detectar a dependência entre sítios. A extensão do LD pode ser influenciada por vários fatores e permite identificar marcadores genéticos que interagem com as variantes causais reais para doenças humanas complexas. É importante a filtragem dos SNP que estão em equilíbrio de ligação para evitar a forte influência de agrupamentos de marcadores SNPs, por exemplo, na análise de Componentes Principais e Análises de Relacionamento entre indivíduos. Utilizamos a medida  $r^2$  para LD, tendo em vista que é a medida mais comumente usada em marcadores bialélicos, onde o quadrado do coeficiente de correlação representa a relação entre as duas variáveis discretas: uma representando o número de um alelo particular no primeiro sítio, e a outra o número do referido alelo no segundo sítio. Para a criação do banco de dados dos dados genéticos do projeto ISA-Nutrição, utilizamos dois métodos de filtros alternativos. Dentre as abordagens para filtragem de marcadores SNP com base no desequilíbrio de ligação, utilizamos o método *prune* (remoção) pela ferramenta PLINK. Nessa abordagem, o algoritmo percorreu todo o genoma em busca de marcadores SNPs próximos em LD, calculando dentro de uma janela de 50bp o valor de correlação genotípica *pairwise*. Se o valor da correlação encontrada for maior que 0,5, esses marcadores foram removidos recursivamente, permanecendo aquele que tivesse o maior MAF. Além disso, o número para mudança do intervalo foi determinado em 5 marcadores SNPs (Prive et al., 2018).

Na segunda filtragem, utilizamos o mesmo comando da ferramenta PLINK ajustando o comando *indep-pairwise*, para um intervalo em 50bp, como número de SNPs para a mudança de intervalo de 5, e o coeficiente de correlação  $r^2$  definido em 0,11, deixando o banco de dados mais “restrito” no número de marcadores (em equilíbrio de ligação).

- **Filtros por cromossomos determinados**

Para a análise de ancestralidade global outros filtros foram aplicados para reduzir potenciais fatores de confundimento nas análises. Um dos casos foi excluir os cromossomos sexuais e do genoma mitocondrial, devido à variabilidade entre os gêneros e a diversidade do genoma mitocondrial que, estruturalmente, é diferente do genoma nuclear e por ser conhecido pelas altas taxas de mutação (Taylor & Turnbull, 2005). Ao final das filtragens realizamos a conversão dos arquivos para o formato binário determinado pelo programa PLINK, (\*.bed), que anota os genótipos, e dois arquivos de texto, (\*.fam), que contém informações sobre os indivíduos, e (\*.bim), com informações sobre os marcadores genéticos. Os resultados do número de marcadores SNPs usados na análise de ancestralidade após esses filtros encontra-se na Tabela 24.

### 3.7. Ancestralidade Global

- **Comparação com os dados do Projeto 1000 Genomas**

A classificação da ancestralidade global pode ser determinada a partir da combinação dos genótipos estudados com os genótipos das populações de um conjunto de dados de referência com indivíduos de etnias conhecidas. Adotamos a combinação com o Projeto 1000 Genomas (The 1000 Genomes Project Consortium, 2015), uma iniciativa que possui um conjunto de dados de referência com uma estrutura populacional de ancestralidade em grande escala continental, mantidos pelo *International Genome Sample Resource (IGSR)*. Utilizamos a combinação com os dados da Fase 3 do Projeto 1000 genomas, com a seleção de 1585 indivíduos de 4 superpopulações (Tabela 5): Africana (AFR), Americana Miscigenada (AMR), Europeia (EUR) e Leste-Asiática (EAS). A concatenação foi realizada a partir de comandos do customizados para o programa PLINK nas versões v1.9 e v2.0 e, ao final da execução, os arquivos foram convertidos nos formatos binários do programa PLINK.

**Tabela 5.** Participantes do Projeto 1000 genomas utilizados na combinação

<b>Sigla</b>	<b>Superpopulação</b>	<b>Indivíduos</b>
AFR	africana	557
AMR	americana	46
EAS	leste-asiático	504
EUR	européia	478

- **SNPRelate**

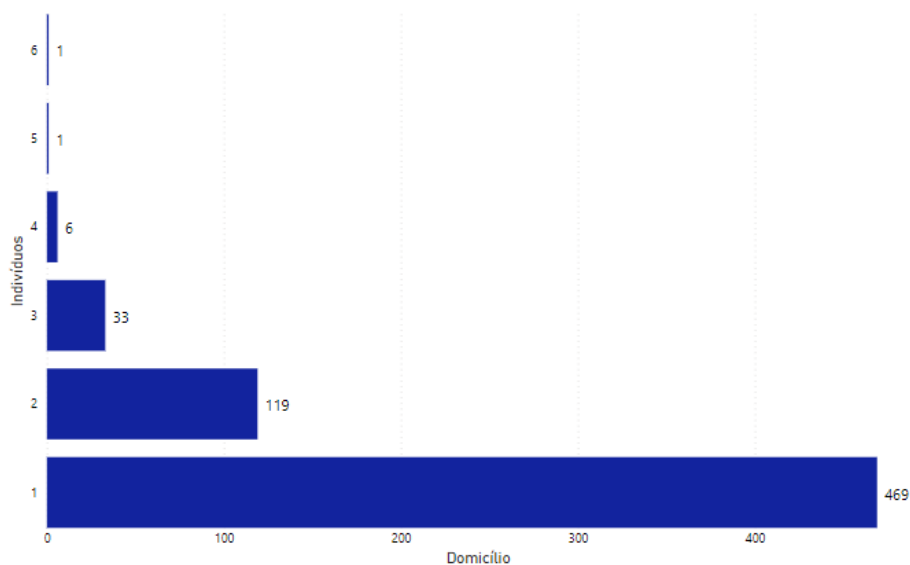
A Análise de Componentes Principais (*Principal Component Analysis* - PCA) é uma técnica de redução de dimensionalidade que permite reduzir o tamanho dos dados pela obtenção de eixos de variação (autovetores com seus correspondentes autovalores), os quais contêm a maior parte da variabilidade dos dados. Os eixos, com as cargas dos autovetores, aplicados às variáveis definem os Componentes Principais (CP). O primeiro CP é a combinação linear das variáveis que explica a maior proporção da variabilidade total, o segundo explica a segunda maior proporção e assim por diante. Gráficos dos autovetores associados (escores dos CPs) com os maiores autovalores são então usados para investigar a estrutura ancestral global da amostra de estudo.

Para o cálculo da PCA utilizamos a biblioteca *SNPRelate* (Zheng et al., 2012) da linguagem R, com o apoio computacional do pacote *gdsfmt*. O pacote *gdsfmt* produz um ambiente otimizado para alto desempenho, fornecendo o uso de memória eficiente, independente de plataforma e gerenciamento de arquivos para dados numéricos de todo o genoma, arquiteturas de computador com processamento *multi-core* (multinúcleo), acelerando os cálculos da Análise de Componentes Principais. Essa otimização é importante ao utilizar a linguagem R, tendo em vista a falta de facilidade da linguagem em produzir customizações para computação paralela.

Para análise dos dados ISA-Nutrição foi desenvolvido um protocolo de processamento na linguagem R para a execução dos comandos dos pacotes *SNPRelate* e *gdsfmt*. Os arquivos no formato binário (\*.bed), (\*.fam) e (\*.bim) foram usados como entrada e foram convertidos pelo comando *snpGDSBED2GDS* no formato (\*.gds). O arquivo (\*.gds) é em formato de arquivo de estrutura de dados genômicos, orientados em matriz, onde um contêiner é utilizado para armazenar os dados de anotação e genótipos de marcadores SNP. Nesse formato, cada *byte* codifica até quatro genótipos de SNP, reduzindo assim o tamanho do arquivo e o tempo de acesso. O formato (\*.gds) oferece suporte ao bloqueio de dados, otimizando que apenas o subconjunto de dados que está sendo processado utilize os recursos de memória. (Zheng et al., 2017) e (Zheng et al., 2012)

O *SNPRelate* através do comando *snpGDSPCA* produz o cálculo otimizado das pontuações (autovetores) dos indivíduos, bem como dos correspondentes autovalores, para as 50 primeiras direções ótimas da PCA. A análise de Componentes Principais Clássica supõe observações independentes. No estudo ISA-Nutrição a coleta de dados foi feita por unidade domiciliar, sendo que em alguns domicílios mais de um indivíduo participou do estudo (Figura 18) o que atribuiu uma possível estrutura de dependência familiar entre os indivíduos. Sob esse cenário, realizamos a análise de PCA para uma subamostra do ISA-Nutrição supostamente formada por indivíduos independentes. Neste caso, um filtro foi aplicado nos dados utilizando o código de domicílio de cada

indivíduo e o parentesco com o responsável pela unidade domiciliar, o que conduziu a 707 indivíduos do total de 841. Com os resultados desta análise, os escores de ancestralidade para o restante da amostra (indivíduos possivelmente dependentes dos demais) são inferidos considerando que estes são uma amostra aleatória dos demais. Esta técnica está implementada no *snpGdsPCA* e foi a opção algorítmica utilizada na análise do ISA-Nutrição.



**Figura 18.** Contagem de Indivíduos por Domicílio.

### 3.8. Copy Number Variation (CNV) e Loss of Heterozygosity (LOH)

Para o presente projeto desenvolvemos um protocolo de detecção e caracterização das chamadas de variações do número de cópias (*Copy Number Variation - CNV*) e das regiões com perda de heterozigosidade (*Loss of Heterozygosity - LOH*) ao longo do genoma, com base no algoritmo *apt-copynumber-axiom-hmm* (*Affymetrics*). O processamento foi aplicado a todos os 841 indivíduos do projeto ISA-Nutrição.

#### 3.8.1. Construção da Referência

Para uma adequada chamada de variantes do tipo CNVs e LOH, é importante a construção de uma referência própria considerando os ruídos experimentais de dentro e fora das placas de processamento laboratorial. Com os dados derivados do protocolo *Best Practices Genotyping Analysis*, foi possível obter duas importantes métricas de variação: MAPD e WavinessSD.



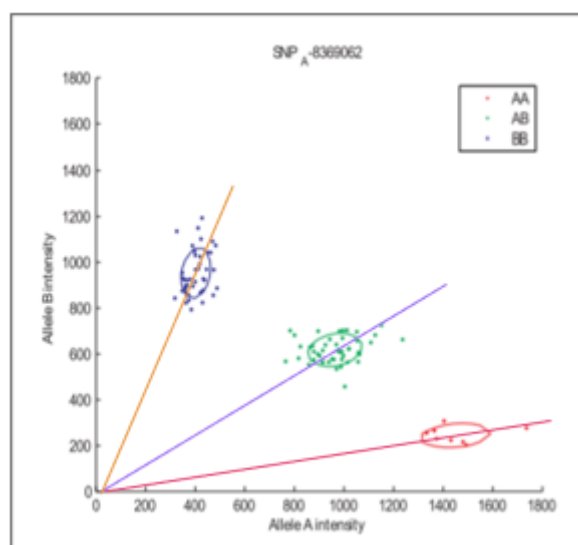
O MAPD (*median absolute pairwise difference*) é uma medida global da variação de todos os conjuntos de sondas em todo o genoma. Ele representa a mediana da distribuição de alterações no  $\log_2$ ratio entre as sondas adjacentes. Uma vez que mede diferenças entre conjuntos de sondas adjacentes, é uma medida de ruído de curto alcance nos dados. Os valores MAPD são os mais sensíveis à referência utilizada. O limite para MAPD é normalmente definido como 0,35. Amostras com valores mais elevados devem ser excluídas do processo de criação de referência.

O *WavinessSD* é uma medida de desvio global de variação das sondas que é insensível a curta variação de alcance e foca na variação de longo alcance. Um alto *WavinessSD* geralmente indica muito ruído nos dados e implica em amostra ou lote de processamento que reduzem a qualidade das chamadas do número de cópias. O limite para *WavinessSD* está definido para 0,1. Amostras com valores mais altos, também devem ser excluídas da construção da referência.

Depois de selecionar as amostras adequadas, estas foram colocadas no protocolo de processamento *Copy Number Reference Creation* do programa *Axiom Analysis Suite* para a criação da referência customizada para a chamada de CNVs e LOHs.

### 3.8.2. Determinação do BAF e LRR para chamada de CNVs e LOH

Para a chamada de CNVs os dados de cada uma das amostras são comparados com uma referência, ou seja, uma análise enriquecida da comparação de cada uma das intensidades (AA, AB e BB) da referência com a amostra analisada. Para tal, as intensidades A e B são transformadas em coordenadas polares R e  $\theta$ , como é possível verificar no exemplo na Figura 19.



$$F(A, B) = (R, \theta)$$

$$R = A + B \quad \text{e} \quad \theta = \frac{\arctang(B/A)}{\pi/2}$$

**Figura 19.** Intensidades dos alelos A e B e equação de determinação dos valores de R e  $\theta$ .

Com os valores das coordenadas R e  $\theta$  é possível calcular as métricas BAF (*B Allele Frequency* – Frequência do Alelo B) e LRR (*Log R Ratio* – Razão Log R), importantes para a determinação do número de cópias de cada segmento e dos trechos de LOH.

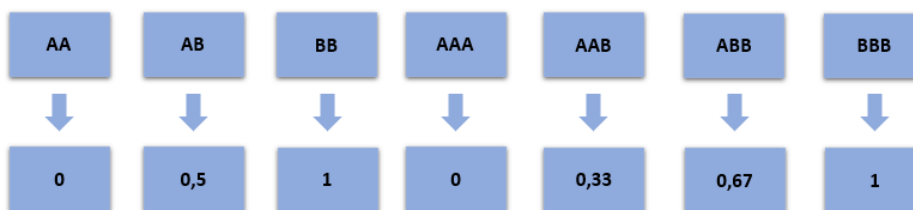
- **Probabilidade do alelo B (BAF)**

Com as informações de  $\theta$  é possível calcular a probabilidade do alelo B (*B Allele Frequency* – BAF), uma medida normalizada da razão de intensidade alélica de dois alelos (A e B). Os valores de BAF determinam a ausência completa de um dos alelos se 0 ou 1 ocorre (por exemplo, AA ou BB) e indica a presença de ambos os alelos quando o valor de BAF é 0,5, no caso, um genótipo AB.

A equação para a determinação do BAF é dada em função de valores de referência  $\theta$ , uma coordenada angular definida em termos de probabilidades genotípicas (K. Wang et al., 2007):

$$BAF = \begin{cases} 0, & \text{se } \theta < \theta_{AA} \\ \frac{0.5(\theta - \theta_{AA})}{(\theta_{AB} - \theta_{AA})}, & \text{se } \theta_{AA} \leq \theta < \theta_{AB} \\ 0.5 + \frac{0.5(\theta - \theta_{AB})}{(\theta_{BB} - \theta_{AB})}, & \text{se } \theta_{AB} \leq \theta < \theta_{BB} \\ 1, & \text{se } \theta \geq \theta_{BB} \end{cases}$$

Em situações em que existe um ganho de cópias, ou seja, a representação de um novo alelo, como por exemplo AAA, AAB, ABB ou BBB, os valores de BAF terão razões sobre qual alelo está presente, que, no caso do exemplo (Figura 20), será, respectivamente, 0, 0,33, 0,67 e 1 (Attiyeh et al., 2009).



**Figura 20.** Genótipos e seus respectivos valores de BAF.

- **Log R Ratio (LRR)**

A partir do valor de R, é calculada a métrica *Log R Ratio* – *LRR* (Razão Log R), uma medida padronizada da intensidade do sinal de cada alelo (A, B) em cada marcador SNP. Neste caso, para os dados de cada marcador é calculado o  $\log_2$  da razão entre o sinal observado e o esperado em duas cópias do genoma. Após a padronização, espera-se que o sinal seja próximo a 0 quando o número de cópias está compatível com o estado normal diploide. Os valores mais altos de LRR podem indicar eventos de duplicação e valores mais baixos podem ser uma evidência de deleções de segmentos cromossômicos (de Araújo Lima & Wang, 2017) (Figura 21). O LRR é dado por:

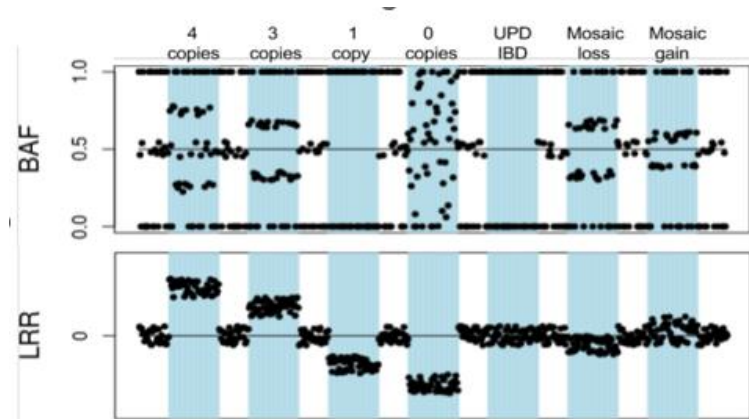
$$LRR = \log_2 \frac{R_{observed\ intensity}}{R_{reference\ intensity}}$$



**Figura 21.** LRR e seus respectivos números de cópias.

### 3.8.3. Chamada de CNV e LOH

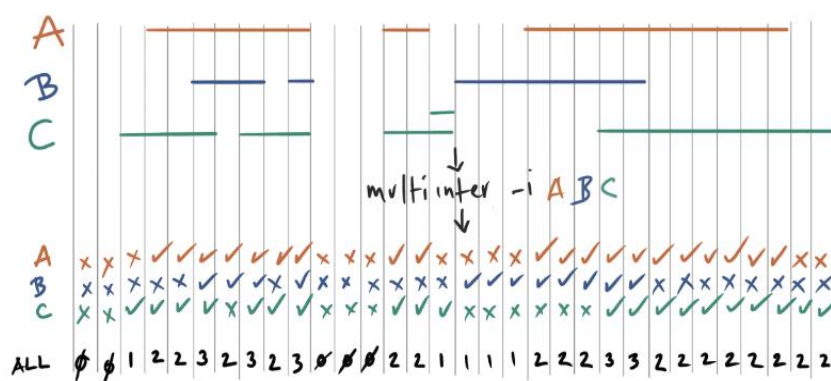
Com a combinação dos valores de LRR e BAF é possível inferir o número de cópias das regiões. Logo, quando ocorre uma deleção, há diminuição dos valores de LRR e uma ausência de heterozigotos nos valores de BAF. Na presença de uma duplicação, há um aumento nos valores de LRR e uma separação do genótipo heterozigoto em dois grupos, em que o número de cópias é utilizado para inferir perdas e ganhos de segmentos cromossômicos. A categorização é realizada pelas classes -2, -1, 0, 1 e 2, em que, o valor 0 é atribuído ao número de cópias normais, os valores negativos para perdas em homozigose ou heterozigose e os valores acima de 0 como ganhos (Figura 22). Na categorização das regiões com LOH, identificamos os marcadores SNPs onde ocorreu uma perda de heterozigosidade, isto é,  $n(Aa)=0$ .



**Figura 22.** LRR e BAF e seus respectivos números de cópias. Adaptado de (Alkan et al., 2011)

No protocolo de detecção de CNVs utilizamos o protocolo *Copy Number Discovery* do programa *Axiom Analysis Suite*, baseado no algoritmo *apt-genotype-axiom* (*Thermo Fisher Scientific*), utilizado para inferir os números de cópias e os estados de perda de heterozigosidade (LOH) com base em um modelo ajustado de Cadeias de Markov Ocultas (Hidden Markov Model – HMM). A HMM é uma técnica estatística que modela um processo de Markov, onde a probabilidade de observar um determinado estado em um determinado ponto de tempo (ou espaço) depende apenas dos estados em pontos anteriores (ocultos). A HMM fornece uma estrutura estatística natural para modelar estruturas de dependência entre números de cópias em marcadores SNPs próximos (Eddy, 2004).

Ao final, foram geradas chamadas separadas para cada indivíduo e convertidas para o formato textual (\*.bed) contendo as colunas cromossomo, posição inicial e posição final, separadas por tipos de CNVs: deleções e duplicações, e LOH. Com protocolo de processamento na linguagem bash, foi utilizado o algoritmo *bedtools multiinter* para encontrar as regiões sobrepostas com pelo menos 1bp, conforme ilustrado na Figura 23 para os indivíduos A, B e C. Ainda, foi adicionada uma quarta coluna com a frequência da região.



**Figura 23.** Exemplo da sobreposição das chamadas, adaptado de (Quinlan & Hall, 2010).

## 4. Resultados e Discussão

### 4.1. Disponibilização dos Bancos de Dados Genômicos

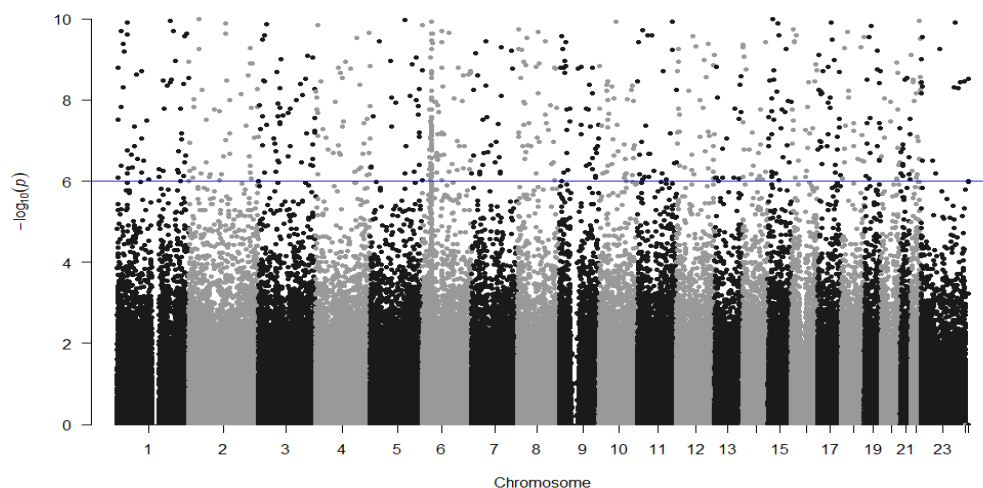
Os bancos de dados genéticos do projeto ISA-Nutrição foram processados nas versões de genoma de referência GRCh37 e GRCh38 do genoma humano. Como já citado, foram disponibilizadas 8 versões de cada montagem, conforme é possível observar na Figura 24, onde varia o valor  $r^2=0,5$  de Desequilíbrio de Ligação, valor de  $r^2=0,11$  de Desequilíbrio de Ligação, e variações dos filtros nos marcadores SNPs.

Bancos Disponibilizados	Nº marcadores
<b>Banco de Dados Estendido de Genótipos dos SNPs total</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,0001, Eq-HW 0,000001.	741.974
<b>Banco de Dados Estendido de Genótipos dos SNPs com rs</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,0001, Eq-HW 0,000001, com informação de rs.	731.392
<b>Banco de Dados de Genótipos de SNPs com rs em Desequilíbrio de Ligação Maior</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,0001, Eq-HW 0,000001, com informação de rs, LD com $r^2 = 0.5$	591.963
<b>Banco de Dados de Genótipos de SNPs com rs em Desequilíbrio de Ligação Menor</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,0001, Eq-HW 0,000001, com informação de rs, LD com $r^2 = 0.11$	315.753
<b>Banco de Dados Estendido de Genótipos dos SNPs total MAF 1%</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,01, Eq-HW 0,000001.	532.873
<b>Banco de Dados Estendido de Genótipos dos SNPs com rs MAF 1%</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,01, Eq-HW 0,000001, com informação de rs.	526.628
<b>Banco de Dados de Genótipos de SNPs com rs em Desequilíbrio de Ligação Maior MAF 1%</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,01, Eq-HW 0,000001, com informação de rs, LD com $r^2 = 0.5$	408.507
<b>Banco de Dados de Genótipos de SNPs com rs em Desequilíbrio de Ligação Menor MAF 1%</b> Recomendados, autossômicos, fora de regiões palindrômicas/HLA, MAF 0,01, Eq-HW 0,000001, com informação de rs, LD com $r^2 = 0.11$	181.477

Figura 24. Bancos Disponibilizados para o Projeto ISA - Nutrição

#### 4.1.1. Gráfico *Manhattan*

Através do gráfico *Manhattan* (Figura 25) observamos a distribuição dos marcadores ao longo de todo o genoma valor p do teste de Equilíbrio de *Hardy-Weinberg* (HWE). Estudos correlacionam os valores altamente desviados do equilíbrio como resultantes de erros de genotipagem (Marees et al., 2018). No gráfico, o valor de corte  $-\log_{10}(p)=10^{-6}$  está indicado, o qual foi adotado durante a filtragem por HWE, de modo a manter o valor menos restrito de marcadores.



**Figura 25.** Gráfico *Manhattan* dos valores p do teste HWE

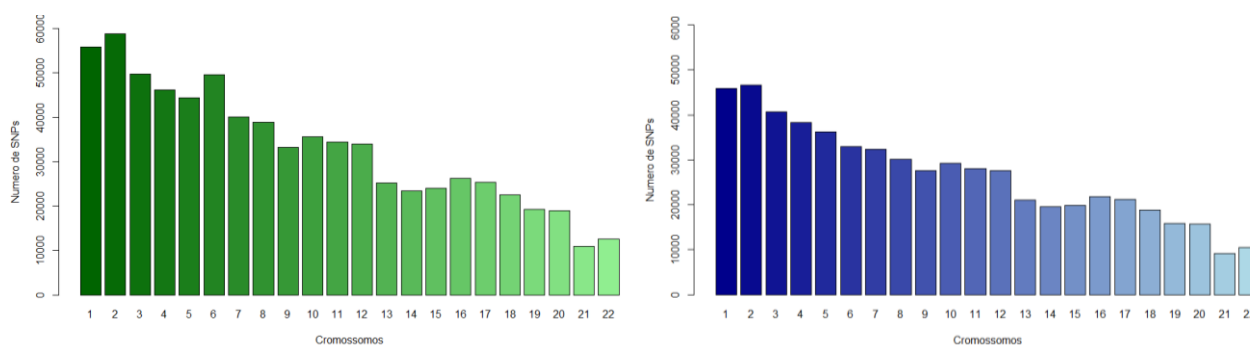
Os valores observados corroboram que a genotipagem das amostras e dos marcadores possuem robustez e boa qualidade. Os valores p obtidos no teste HWE estão dentro do padrão obtido em estudos de associação de genoma (GWAS), em que, utilizando um ponto de corte de HWE < 0,01 tivemos 14.995 SNPs, com HWE < 0,001 tivemos 3.930 SNPs, HWE < 0,00001 tivemos 1.001 SNPs e HWE < 0,000001 SNPs identificados 731 SNPs em desequilíbrio de HWE. Para a Menor Frequência do Alelo (MAF), com MAF < 0,05 tivemos 473.262 marcadores, com MAF < 0,01 tivemos 277.470 marcadores e com MAF < 0,0001 tivemos 72.634 marcadores (Tabela 6). Todos estes filtros descritos estão restritos aos cromossomos autossômicos, e os resultados de marcadores excluídos do banco de dados foram apropriadamente documentados. Para a análise de ancestralidade dos dados do ISA-Nutrição foi adotado o filtro HWE < 0,000001 e MAF < 0,0001, além de alguns outros como descrito na tabela 6. Ao final de cada etapa de filtragem restaram 557.426 marcadores.

**Tabela 6.** Número de marcadores SNPs excluídos por filtros e os valores restantes durante o processamento.

Filtro	SNPs excluídos	SNPs restantes
Filtro Recomendados	47.568	873.177
SNPs com identificador "rs_id"	35.084	838.093
Filtro autossômicos	32.875	805.218
MAF (0,0001)	72.634	732.584
HWE (0,000001)	731	731.853
HLA-Palindromics	461	731.392
high-LD-regions-hg19-GRCh37	20.305	711.087
LD r2>0.5 in a 50b window, shift step=5	119.124	591.963

Comuns com 1000g	4.899	587.064
Filtro Multiallelic SNPs	29.638	557.426

Avaliamos a distribuição dos marcadores por cromossomo e comparamos o comportamento antes da etapa de remoção pelo método *pruning* para desequilíbrio de ligação (LD), com 731.392 marcadores, e depois da filtragem, com 591.963 marcadores (Figura 26).



**Figura 26.** Na esquerda (em verde) o número de SNPs ao longo dos cromossomos antes da etapa de remoção por *pruning* e na direita (em azul) o número de SNPs para os cromossomos depois da etapa de remoção por *pruning*.

## 4.2. Análise de Ancestralidade Global - Análise de Componentes Principais (PCA)

Como método para a avaliação dos grupos ancestrais utilizamos a Análise de Componentes Principais (PCA) para o cálculo de um mapa de ancestralidade, através do comando *snpGdsPCA* do *SNPRelate*. Dos 841 integrantes do estudo, 134 foram separados da primeira etapa de processamento, pois foram identificados como apresentando algum grau de parentesco com outros indivíduos do domicílio, assim restaram 707 participantes. O critério de escolha adotado foi de manter apenas 1 indivíduo quando houve parentesco. Logo, a matriz de genótipos, denominada como G, com dimensão NxP, onde N = 2292 indivíduos (Projeto ISA-Nutrição + Projeto 1000 genomas) e P = 555.064 marcadores foi utilizada na análise. Estatísticas resumo dos escores dos 4 primeiros CP estão apresentadas na Tabela 7.

**Tabela 7.** Estatísticas resumo dos 4 primeiros Componentes Principais (escores) encontrados nos dados do Projeto ISA-Nutrição em conjunto com os indivíduos do Projeto 1000 genomas.

CP1	CP2	CP3	CP4
Min. :-0.02090	Min. :-0.0259207	Min. :-0.1468540	Min. :-8.269e-02
1st Qu.: -0.01432	1st Qu.: -0.0209229	1st Qu.: 0.0005936	1st Qu.: -3.476e-03
Median : -0.01139	Median : 0.0002209	Median : 0.0030559	Median : -8.868e-05
Mean : 0.00000	Mean : 0.0000000	Mean : 0.0000000	Mean : 0.000e+00
3rd Qu.: 0.02135	3rd Qu.: 0.0078796	3rd Qu.: 0.0095382	3rd Qu.: 2.603e-03
Max. : 0.03605	Max. : 0.0347819	Max. : 0.0150327	Max. : 6.804e-02

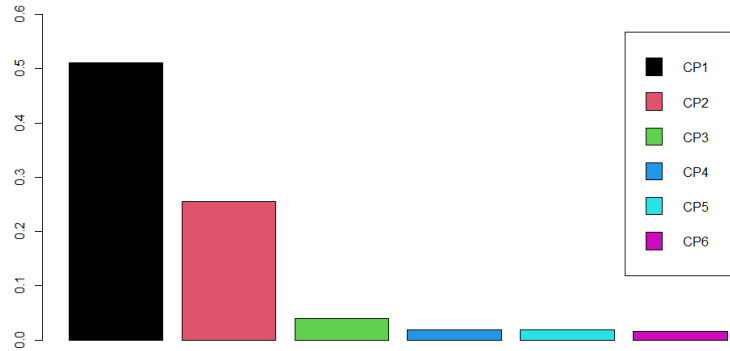
Após a etapa de cálculo dos CPs, através dos comandos *snpGdsPCASNPLoading* e *snpGdsPCASampLoading* em um protocolo desenvolvido na linguagem R, as 134 amostras excluídas da análise anterior, foram agrupadas aos dados, inferindo o cálculo dos autovetores dos SNPs com base nos valores da PCA baseada nas 707 amostras + amostras do Projeto 1000 genomas.

**Tabela 8.** Número de SNPs excluídos por filtros e os valores restantes para o processamento da ancestralidade global dos 841 indivíduos.

Filtro	SNPs excluídos	SNPs restantes
Filtro Recomendados	47.568	873.177
SNPs com rs	35.084	838.093
Filtro autossômicos	32.875	805.218
MAF (0,0001)	74.789	730.429
HWE (0,000001)	590	729.839
HLA-Palindromics	459	729.380
high-LD-regions-hg19-GRCh37	20.263	709.117
LD r <sup>2</sup> >0.5 in a 50b window, shift step=5	119.902	589.215
Comuns com Projeto 1000 genomas	4.583	584.632
Filtro Multiallelic SNPs	29.568	555.064

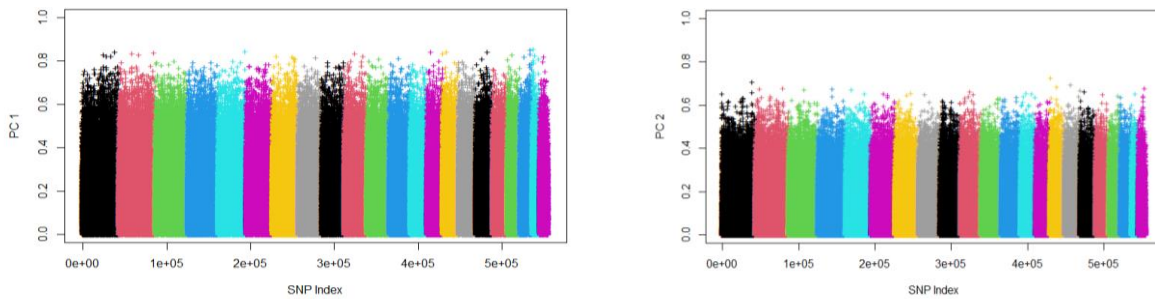
Nos dados concatenados entre o projeto ISA-Nutrição e Projeto 1000 genomas, os dois primeiros Componentes Principais somados representaram cerca de 76,5% do mapa de ancestralidade (Figura 27), tornando forte a representação dos marcadores com a ancestralidade.



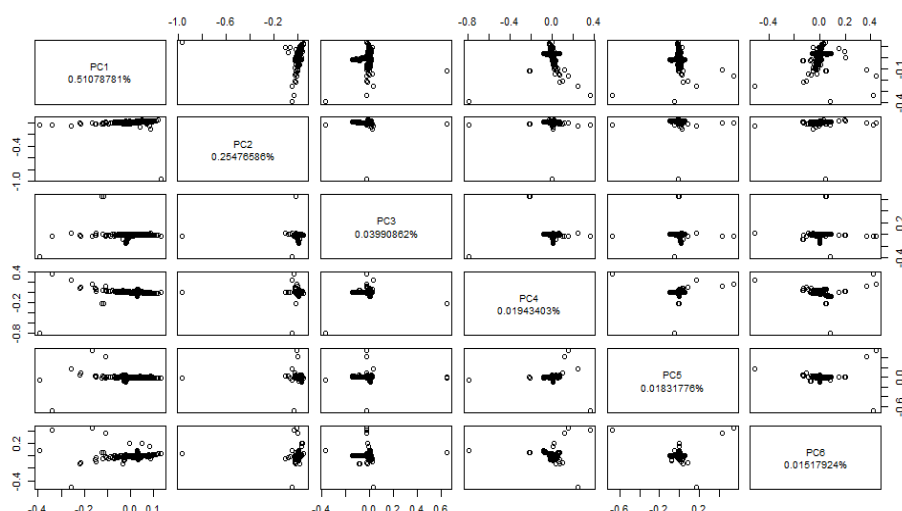


**Figura 27.** Porcentagem da explicação dos 6 primeiros componentes principais com ancestralidade.

Para entender a representação dos CP1 e CP2 para cada cromossomo, calculamos a correlação com os marcadores ao longo do genoma, através do comando *snpGdsPCACorr* da biblioteca *SNPRelate* (Figura 28). De maneira geral, o padrão de correlação indica correlações de altas a moderadas dos CP1 e CP2 para todos os marcadores avaliados nos 22 autossomos. A Figura 29 mostra a dispersão dos escores para os 6 primeiros CP.



**Figura 28.** Na esquerda a correlação do CP1 (51,1% de explicação) e na direita a correlação do CP2 (25.4%) com os genótipos ao longo do genoma.

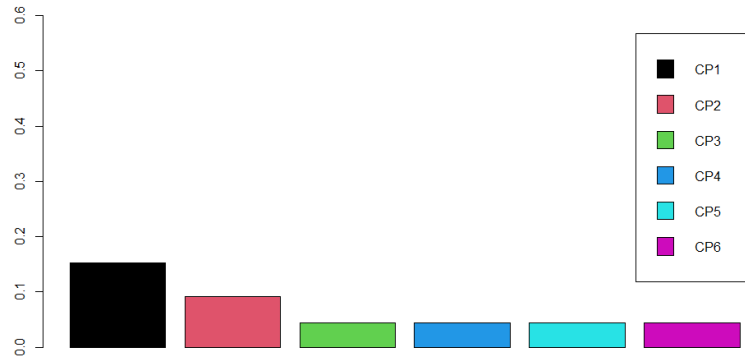


**Figura 29.** Relacionamento dos 6 primeiros CP da análise do Projeto *ISA-Nutrição* + Projeto 1000 genomas.

Calculamos também os Componentes Principais para os dados apenas do Projeto *ISA-Nutrição*, onde a matriz  $G$  com dimensão  $N \times P$ , teve  $N = 707$  indivíduos e  $P = 589.215$  de marcadores utilizado na análise. Com base no resultado desta análise, o restante das amostras (134 indivíduos) teve seus escores inferidos. Observando os dois primeiros Componentes Principais (CP), encontramos uma representação de aproximadamente 24,4% (Tabela 9 e Figura 30). Este resultado mostra o impacto de enriquecer a amostra do estudo com os dados do Projeto 1000 genomas na análise de ancestralidade, o que permite que as ancestralidades dos participantes do *ISA-Nutrição* migrem para suas correspondentes origens ancestrais.

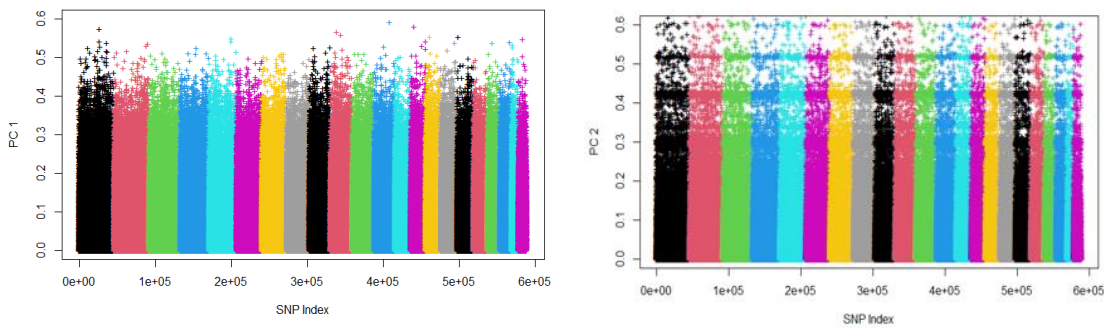
**Tabela 9.** Estatísticas resumo dos 4 primeiros Componentes Principais encontrados nos dados do Projeto *ISA-Nutrição*.

CP1	CP2	CP3	CP4
Min. :-0.145767	Min. :-0.3265595	Min. :-0.5309358	Min. :-0.4436725
1st Qu.: -0.018625	1st Qu.: 0.0002737	1st Qu.: -0.0014931	1st Qu.: -0.0003791
Median : 0.004968	Median : 0.0043597	Median : -0.0006098	Median : 0.0000392
Mean : 0.000000	Mean : 0.0000000	Mean : 0.0000000	Mean : 0.0000000
3rd Qu.: 0.025415	3rd Qu.: 0.0090810	3rd Qu.: 0.0004907	3rd Qu.: 0.0004269
Max. : 0.045578	Max. : 0.0181222	Max. : 0.2173020	Max. : 0.7982912

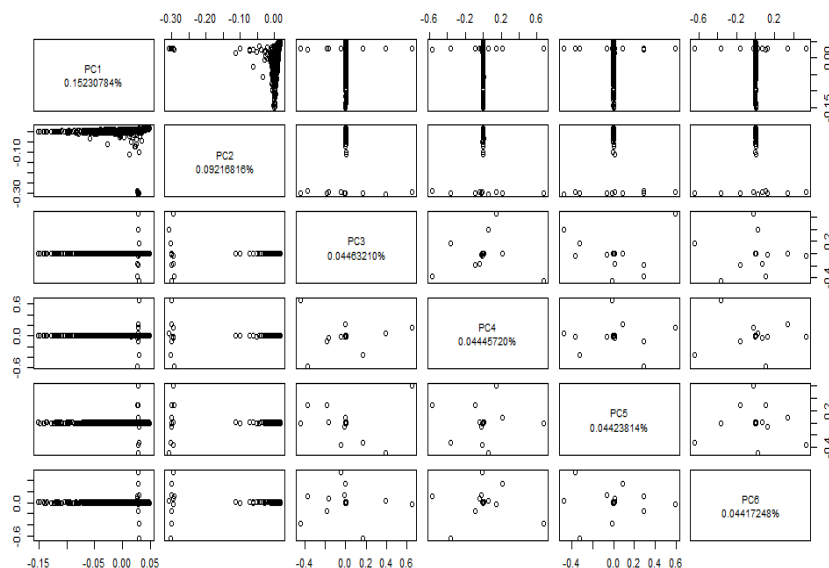


**Figura 30.** Porcentagem da explicação dos 6 primeiros componentes principais com ancestralidade.

Verificamos a correlação do CP1 e CP2 com os marcadores ao longo do genoma, para os dados apenas do Projeto ISA-Nutrição (Figura 31), em que, comparado com os dados enriquecidos pelo Projeto 1000 genomas, valores mais baixos de correlação são encontrados. A Figura 31 mostra a dispersão dos escores para os 6 primeiros CP.



**Figura 31.** Na esquerda a correlação do CP1 (15.2% de explicação) e na direita a correlação do CP2 (9.2%) com os genótipos ao longo do genoma.

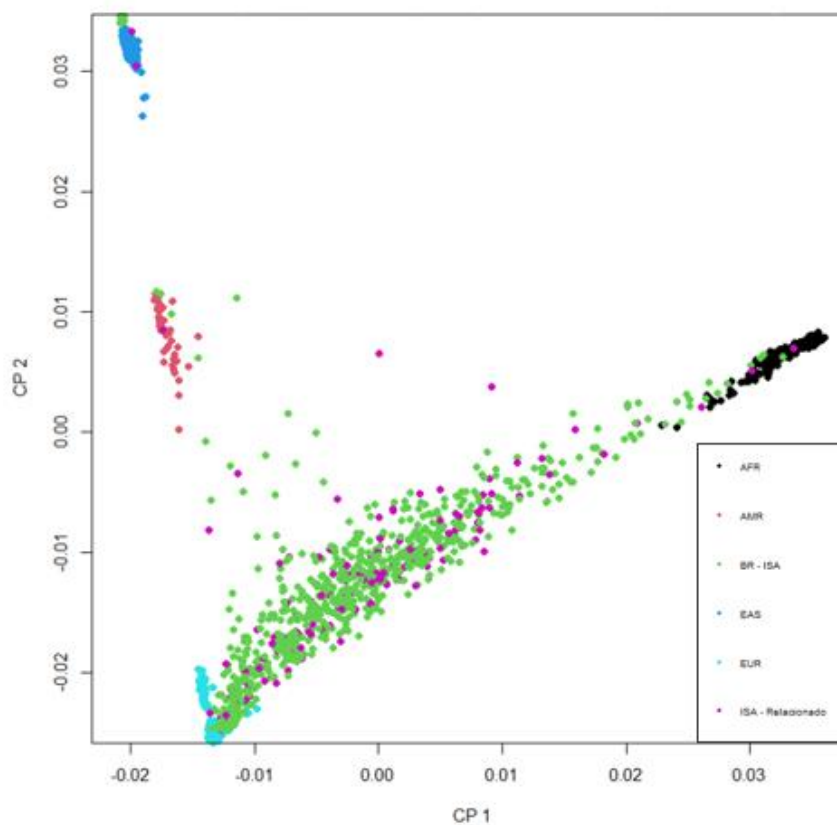
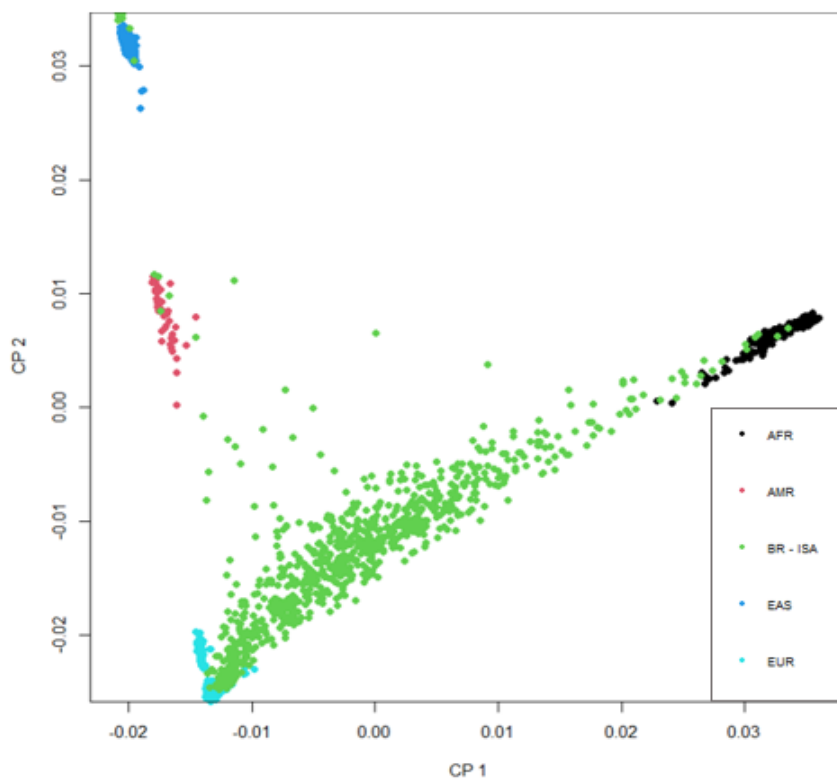


**Figura 32.** Relacionamento dos 6 primeiros CP da análise do Projeto ISA-Nutrição.

É notável que, ao incluir nas análises das diferentes etnias representadas no Projeto 1000 genomas, os CPs tornaram-se mais correlacionados com a explicação da ancestralidade em comparação aos valores obtidos com as análises apenas dos indivíduos do Projeto ISA-Nutrição.

### 4.3. Gráficos de Ancestralidade Global

Com o cálculo de PCA utilizamos os valores de CP1 e CP2 para verificar graficamente o mapa de ancestralidade, ou seja, a disposição entre as diferentes etnias do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição. No gráfico (Figura 33), como referência estão as superpopulações AFR = Africana, AMR = Americana, EAS = Leste-Asiática e EUR = Europeia, que revelam a alta miscigenação do grupo de brasileiros, BR = Indivíduos do Projeto ISA-Nutrição.

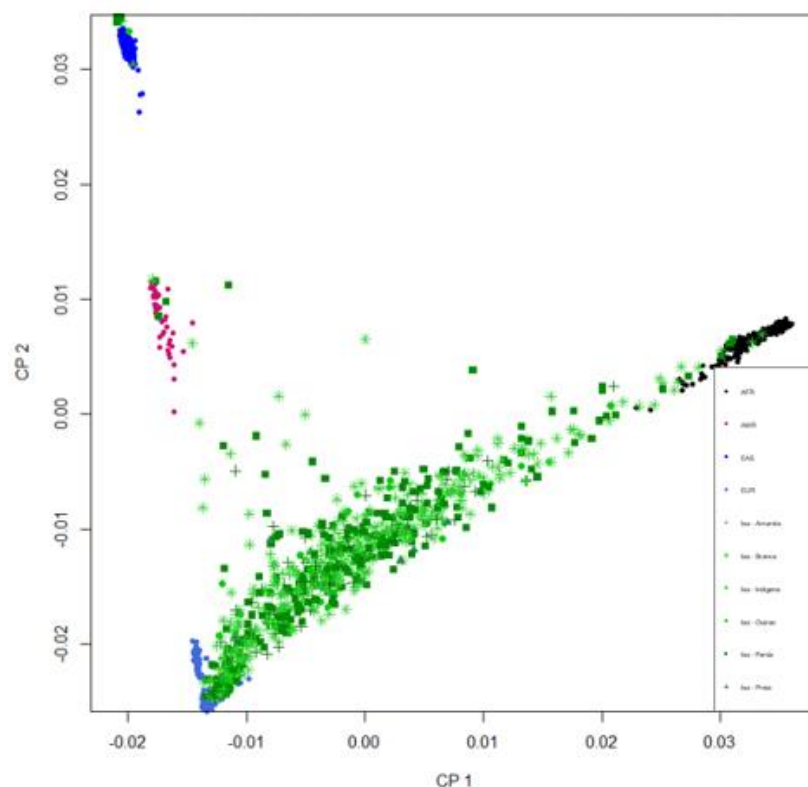


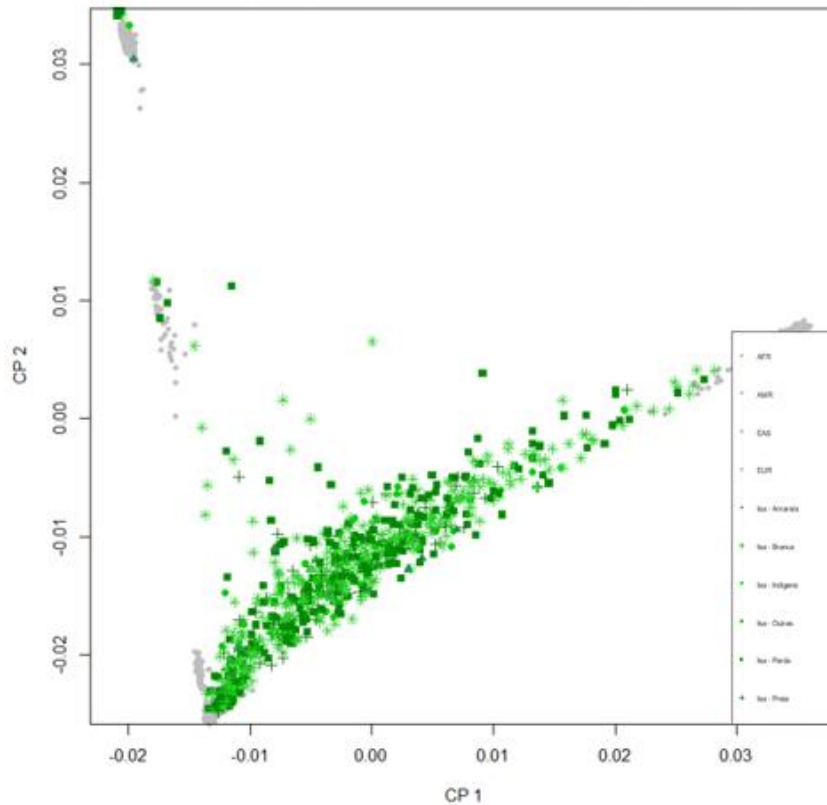
**Figura 33.** (1) Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição pelos CP1 e CP2. (2) Representação das 4 superpopulações do Projeto 1000

genomas e dos indivíduos do Projeto ISA-Nutrição pelos CP1 e CP2 com destaque nos 134 indivíduos relacionados, excluídos na primeira etapa do cálculo.

Pelos resultados obtidos é possível notar o alto grau de miscigenação das amostras do Projeto ISA-Nutrição, em que os indivíduos mostram grande dispersão interna com similaridades aos valores de CPs obtidos em quase todas as etnias avaliadas. Os resultados corroboram com outros estudos caracterizando a alta miscigenação da população brasileira (Giolo et al., 2012; Santos et al., 2016; Secolin et al., 2019)

Para uma caracterização da ancestralidade global dos indivíduos utilizamos as categorizações de declaração de raça e local de nascimento para estratificar a amostra do projeto ISA-Nutrição em subgrupos de acordo com cada classificação. Para a declaração de raça utilizamos a divisão de acordo com a categoria utilizada em conformidade com a normatização do Instituto Brasileiro de Geografia e Estatística (IBGE): Isa – Amarela, Isa – Branca, Isa – Indígena, Isa – Parda, Isa – Preta, e Isa – Outras, para os valores não informados (Figura 34).



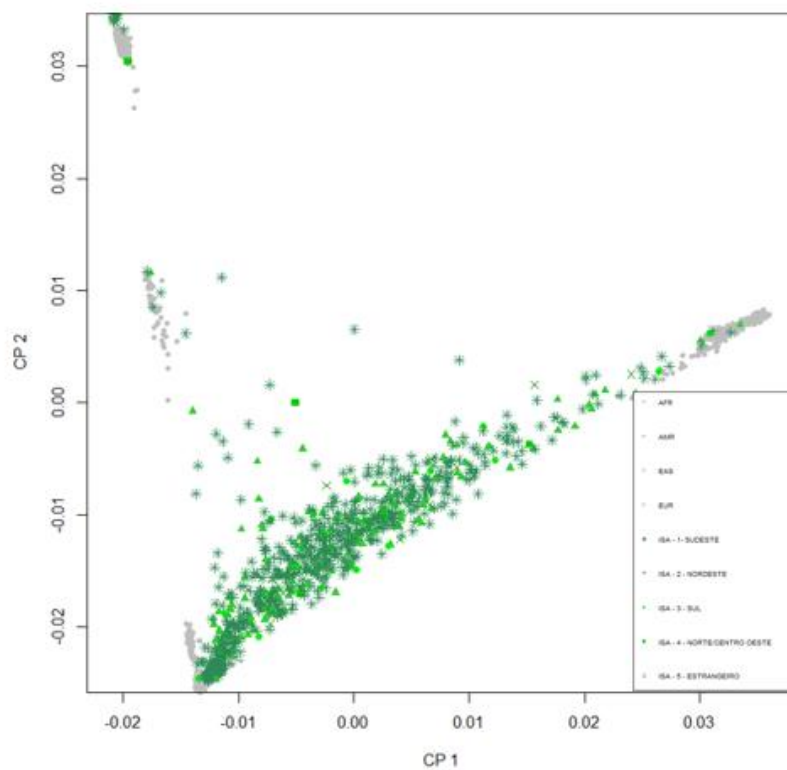
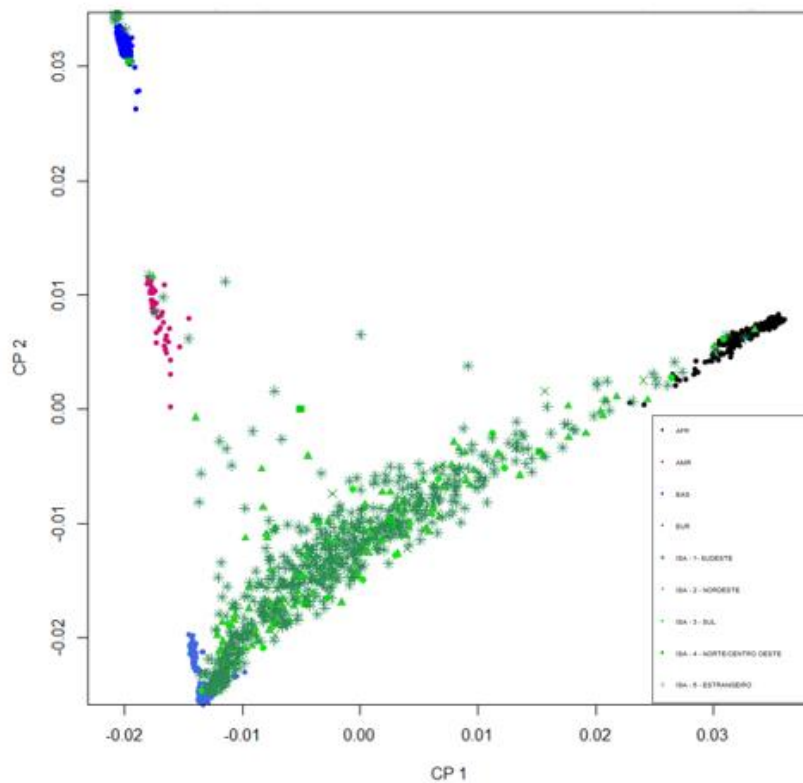


**Figura 34.** Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição categorizados por raça autodeclarada pelos CP1 e CP2. Na primeira figura com a informação das superpopulações e na segundo o destaque nas subcategorias do projeto ISA-Nutrição.

Ao observar o gráfico é possível notar que nem sempre a raça autodeclarada é uma informação concordante com o escore ancestral global, o que é esperado principalmente em uma população heterogênea como a brasileira. Essa informação é de extrema importância em estudos populacionais, pois usando apenas a referência de raça autodeclarada para a categorização dos indivíduos, o estudo por encontrar fontes de viés, por apenas selecionar as características físicas como determinantes de etnia e levar a conclusões equivocadas (Mersha & Abebe, 2015).

A cidade de São Paulo tem uma alta representatividade de etnias presentes na formação da população brasileira pela alta taxa de migração de indivíduos, que, no caso, pode ocorrer de outros estados em busca de oportunidades de emprego, bem como de outros países por busca de refúgio ou novas oportunidades de vida. Desse modo, ao conhecer a naturalidade dos participantes, espera-se que, indiretamente, podemos conhecer o papel de outras regiões na migração interna e correlacionar com as possíveis superpopulações do Projeto 1000 genomas. Os indivíduos foram separados em 5 grandes categorizações pelo local de nascimento informado: Isa – Sudeste, Isa – Nordeste, Isa – Sul, Isa – Norte/Centro Oeste e Isa – Estrangeiro. Vale ressaltar, que durante a

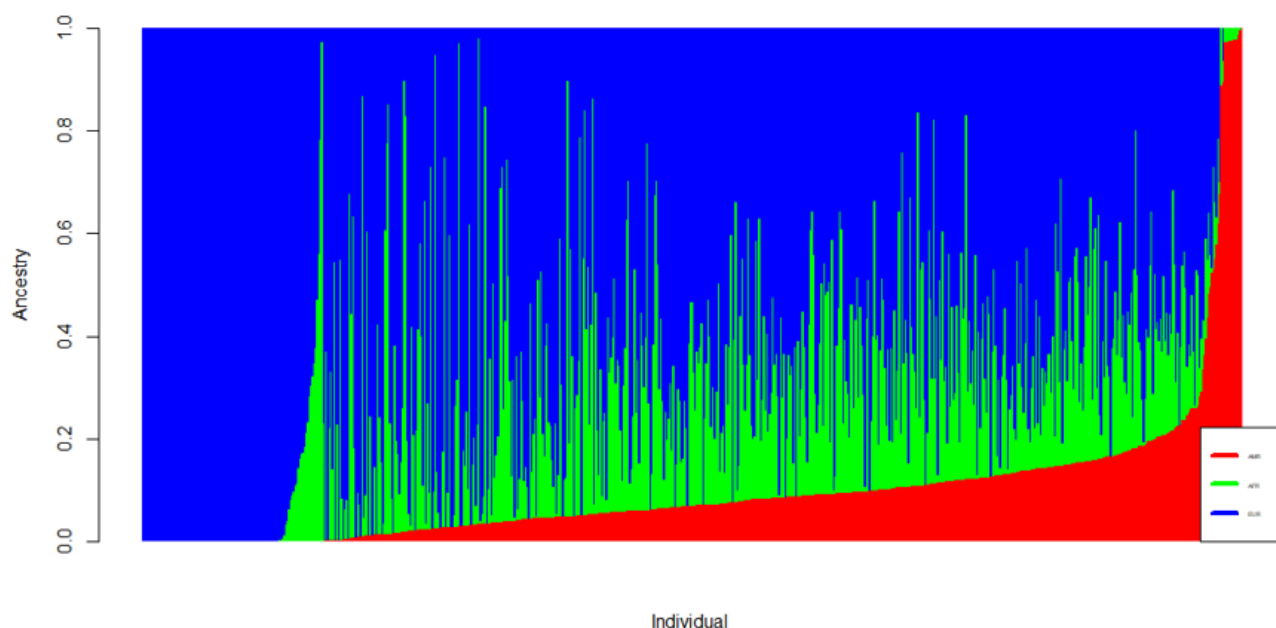
entrevista não foi informado o país de nascimento dos participantes declarados estrangeiros, sendo assim, mantivemos em uma única categorização (Figura 35).





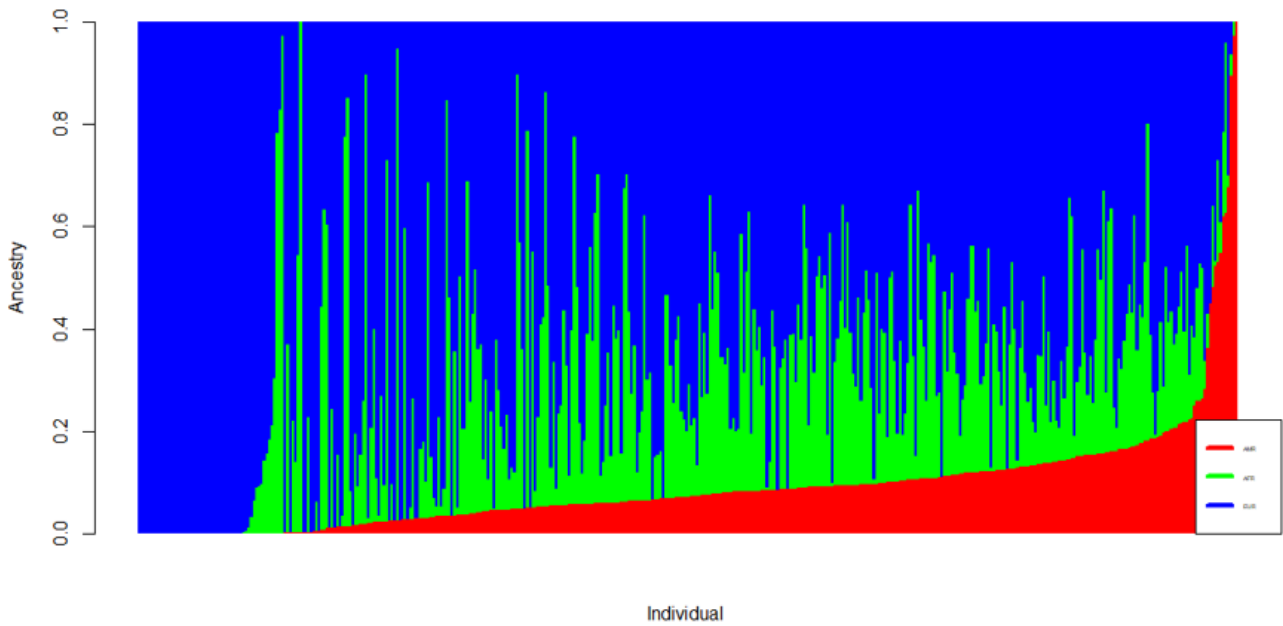
**Figura 35.** Representação das 4 superpopulações do Projeto 1000 genomas e dos indivíduos do Projeto ISA-Nutrição categorizados pelo local de nascimento pelos CP1 e CP2. Na primeira figura com a informação das superpopulações e na segunda o destaque nas subcategorias do projeto ISA-Nutrição.

Devido à alta miscigenação da população brasileira utilizamos recursos para conhecer as proporções ancestrais de cada indivíduo do projeto ISA-Nutrição. Com o comando *snpGDSAdmixProp* do programa *SNPRelate* foram calculadas as proporções ancestrais com base nos valores obtidos durante a análise de ancestralidade via Análise Componentes Principais. Inicialmente, foram utilizadas as 3 principais etnias: AFR = Africana, AMR = Americana e EUR = Europeia, e calculada a matriz de proporções P com NxP, onde N = 841 indivíduos do ISA-Nutrição e P = 3 superpopulações (Figura 36). Apesar da predominância da influência ancestral europeia, é marcante a heterogeneidade e o papel dos dois outros grupos ancestrais (africanos seguidos de ameríndios).



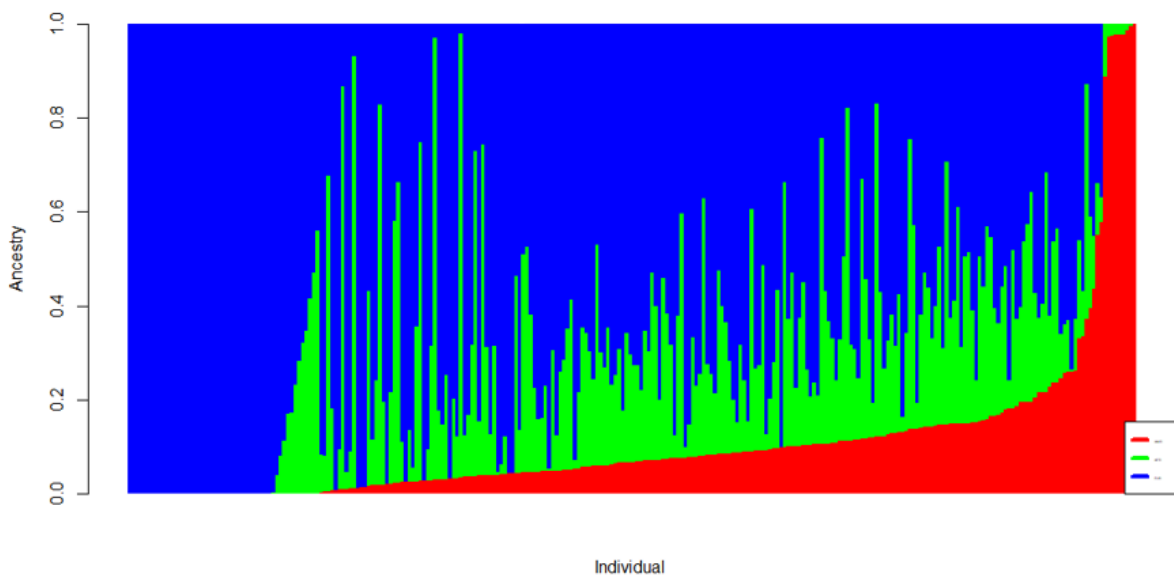
**Figura 36.** Proporções de ancestralidade Individual das amostras do Projeto ISA-Nutrição.

Utilizando as informações de raça autodeclarada dos indivíduos do ISA-Nutrição, a informação das proporções ancestrais contidas na matriz P foi estratificada em subgrupos determinados pela raça mencionada. A Figura 37 mostra a distribuição dos indivíduos autodeclarados brancos de acordo com suas proporções ancestrais das 3 etnias avaliadas (AFR, AMR e EUR).



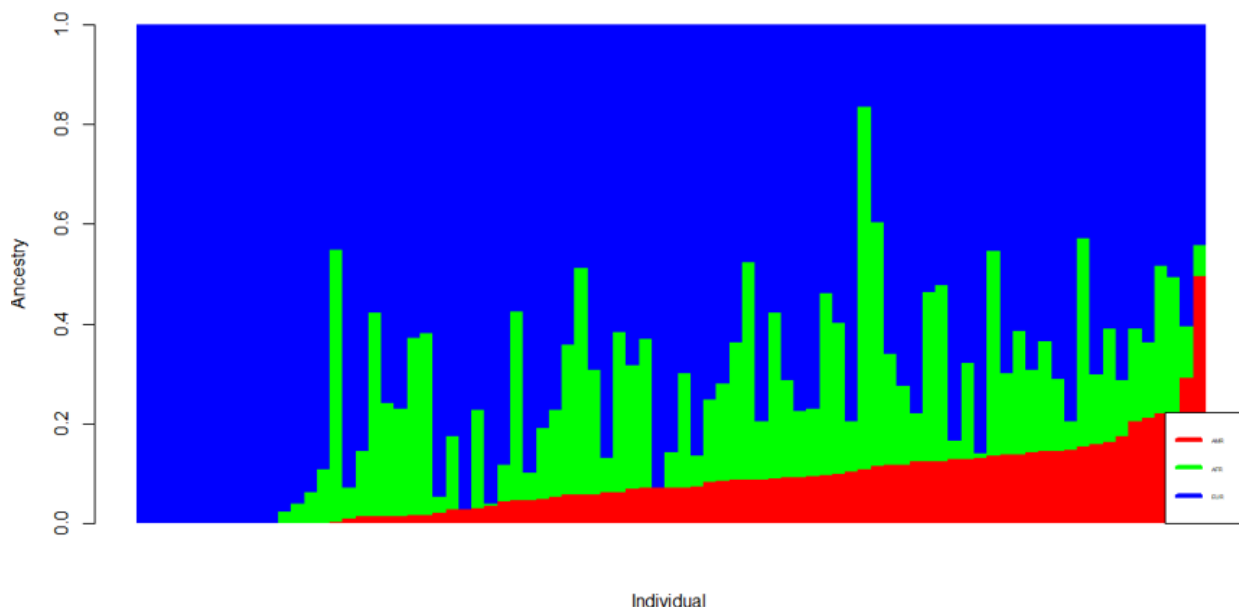
**Figura 37.** Proporções de Ancestralidade Individual de autodeclarados brancos.

Observando a Figura 37 é possível perceber que, apesar da autodeclaração dos indivíduos ser da raça branca, alguns praticamente não possuem proporções da etnia europeia. Também foram calculadas as proporções de ancestralidade individual para os indivíduos autodeclarados pardos e autodeclarados pretos, extraído da matriz P o valor das proporções e combinado com a categoria informada (Figuras 38 e 39).



**Figura 38.** Proporções de Ancestralidade Individual de autodeclarados pardos.

Os indivíduos autodeclarados pardos possuem uma representação bastante miscigenada das 3 etnias comparada aos outros grupos de indivíduos estudados (brancos e pretos). Essas informações são coerentes com a definição do termo pardo, que, segundo o IBGE, é uma raça derivada de uma mistura entre outras raças. Contudo, alguns indivíduos, apesar de se autodeclararem pardos, possuem uma alta representação de uma ou duas etnias (Figura 38).

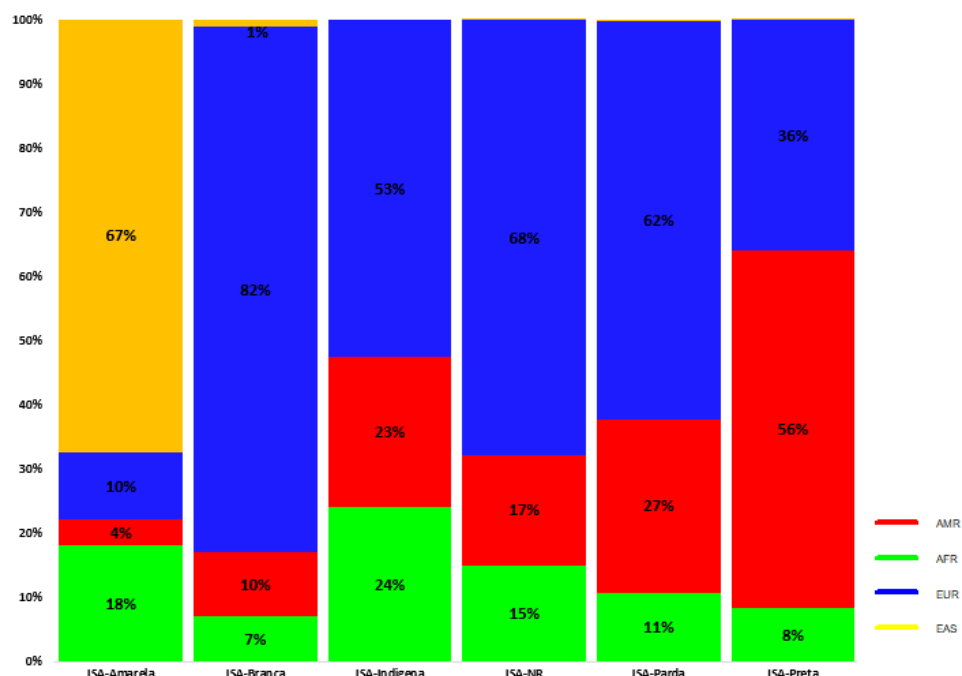


**Figura 39.** Proporções de Ancestralidade Individual de autodeclarados pardos.

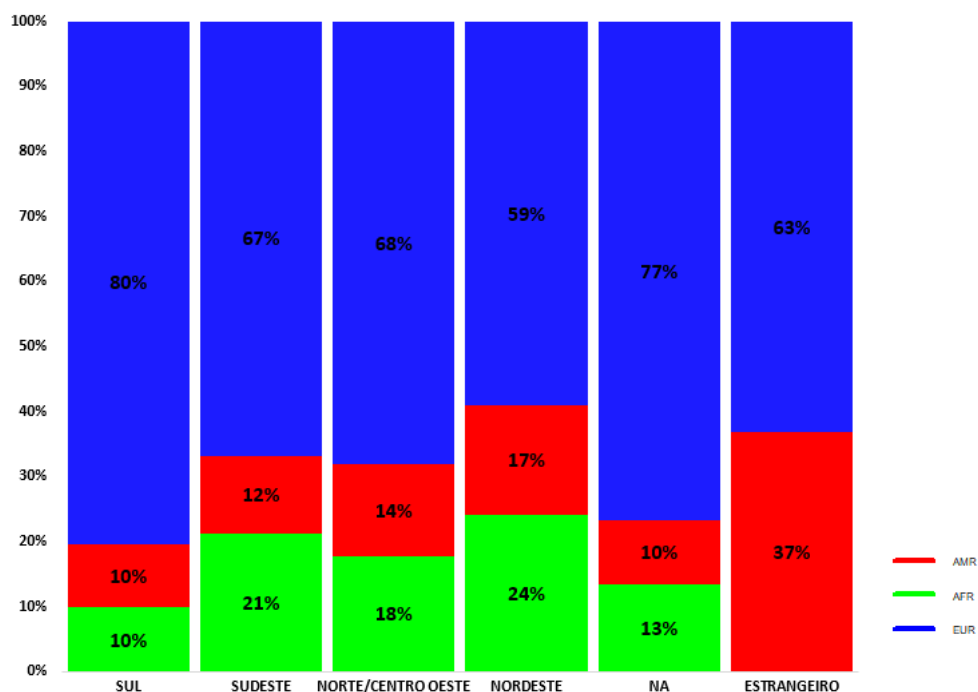
O número de indivíduos autodeclarados pretos foi menor no projeto ISA-Nutrição em comparação com os grupos brancos e pardos. Na análise, os participantes autodeclarados pretos possuem uma boa representação da etnia africana, mas também mostram o padrão de miscigenados e com alta representação da etnia europeia (Figura 39).

A superpopulação asiática EAS = Leste-asiática, teve pouca representatividade nas análises. Dentro do Projeto ISA-Nutrição foram 13 indivíduos autodeclarados amarelos, compondo apenas 1,6% de todo o banco de dados. Contudo, alguns indivíduos autodeclarados amarelos apresentaram curiosamente um alto valor de ancestralidade americana ou nativa. Devido aos vieses apresentados pela autodeclaração da raça e alta miscigenação da população brasileira, é esperado imprecisão nessa informação. Ainda, esse padrão pode ser consequência da correlação dos povos nativos americanos com as populações asiáticas que, possivelmente, deram origem à povoação do continente americano, conforme relatos de estudos paleontológicos (Moreno-Mayar et al., 2018).

A Figura 40 mostra o perfil médio das proporções de ancestralidade de quatro origens ancestrais (AMR, AFR, EUR e EAS) de acordo com a raça autodeclarada e a Figura 41 mostra o perfil médio das proporções de ancestralidade das origens AMR, AFR e EUR para o local de nascimento.



**Figura 40.** Proporção média de ancestralidade por raça autodeclarada. Para cada população tivemos: ISA-Amarela=13, ISA-Branca=422, ISA-Indígena=2, ISA-NR=11, ISA-Parda=309, ISA-Preta=84 (Total 841).

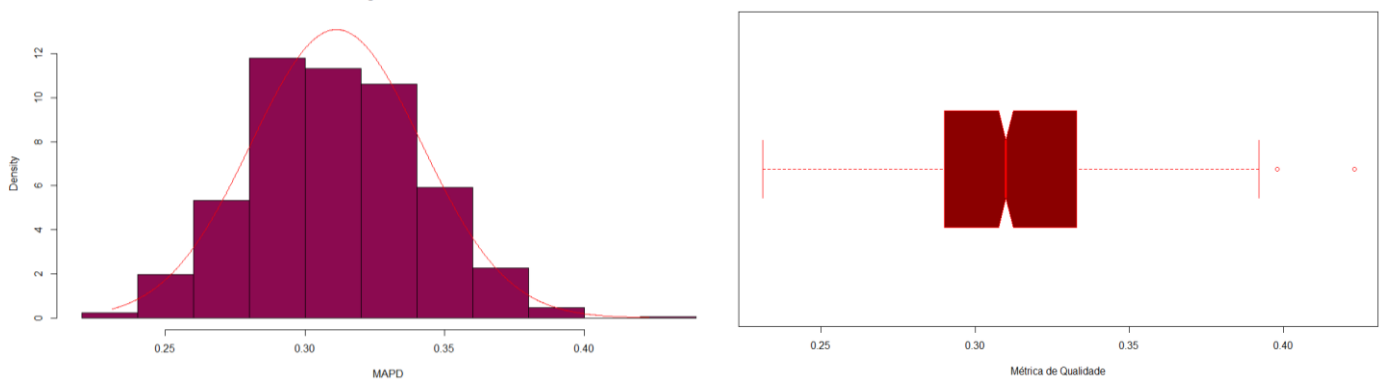


**Figura 41.** Proporção Média de Ancestralidade por Região. Para cada população tivemos: NORDESTE=190, NORTE/CENTRO OESTE=5, ESTRANGEIRO=26, NA=12, SUDESTE=584 e SUL=24 (Total 841).

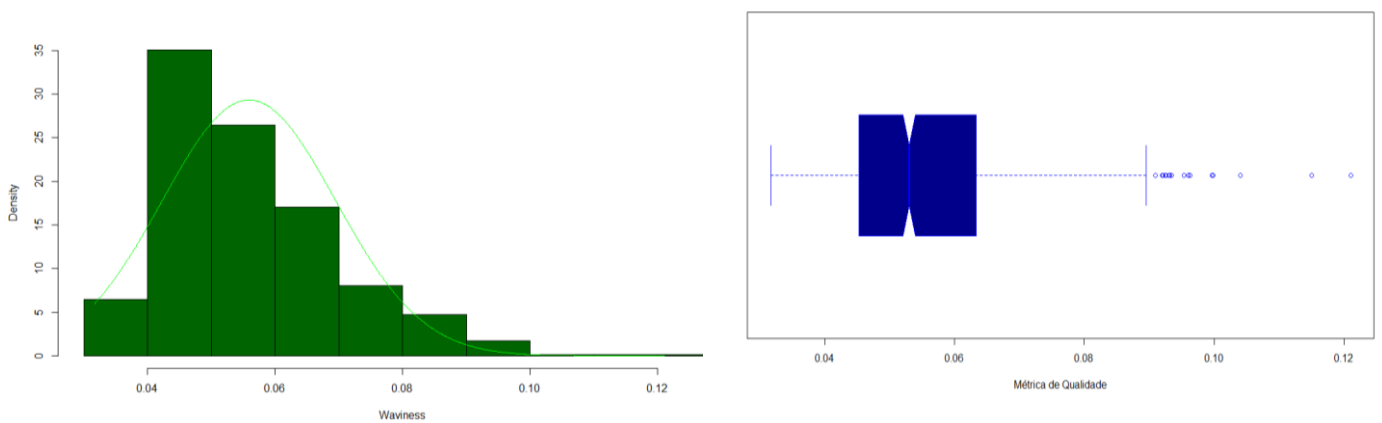
## 4.2. Copy Number Variation (CNV) e Loss of Heterozygosity (LOH)

### 4.2.1. Criação da Referência Personalizada

Das 841 amostras do projeto, nem todas possuíam os requisitos mínimos de qualidade para a construção da referência para as chamadas de CNV e LOH. Usamos como padrão os valores das métricas *MAPD* e *WavinessSD* (descritas na seção de métodos) usando uma execução do protocolo de construção de referência padrão enviada pelo fabricante *Thermo Fisher* e excluimos as amostras da construção da referência personalizada.

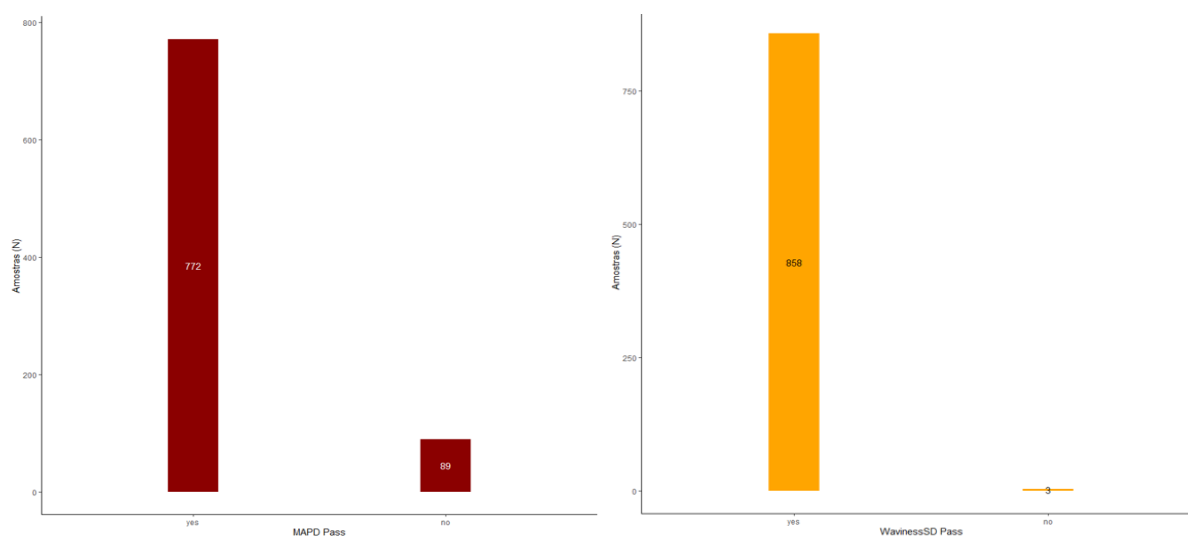


**Figura 42.** (1) Histograma e (2) *Boxplot* (Diagrama de Caixa) mostrando os valores de MAPD para as 841 amostras do Projeto. Ponto de corte recomendado: excluir amostras com valores acima de 0,35.



**Figura 43.** (1) Histograma e (2) *Boxplot* (Diagrama de Caixa) mostrando os valores de *WavinessSD* para as 841 amostras do Projeto. Ponto de corte recomendado: excluir valores acima de 0,10.

Sendo assim, ao final das avaliações, foram excluídas 89 amostras por *MAPD/WavinessSD* e a referência foi construída com o total de 772 amostras, incluindo as amostras de controle processadas nas mesmas placas.



**Figura 44.** Distribuição das amostras do ISA-Nutrição aprovadas e reprovadas na comparação com filtros das métricas de (1) MAPD e (2) *WavinessSD*.

#### 4.2.2. Chamadas de CNV e LOH

- **Filtro de Qualidade**

Com a referência criada, foi iniciado o processamento do protocolo de chamadas de CNV e LOH a partir do programa *Copy Number Discovery* do *Axiom Analysis Suite*, baseado no algoritmo *apt-copynumber-axiom-hmm*. Ao final da execução uma etapa de filtragem foi realizada para eliminar chamadas com tamanhos fora do limite de detecção do algoritmo, conforme instruções do fabricante (Figura 45).

Métrica	Threshold
MAPD	$\leq 0.35$
Waviness SD	$\leq 0.1$

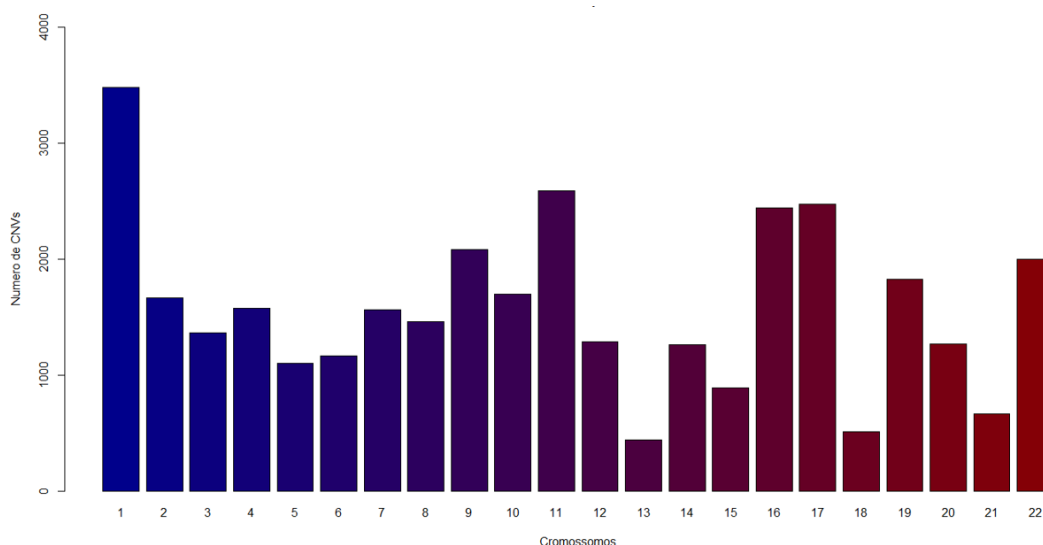
CNV Region QC Thresholds	Threshold
seg-min-bases-CN-1 ou mais	$\geq 50000$
seg-min-bases-CN-0	$\geq 25000$
seg-min-probesets-CN-1 ou mais	$\geq 50$
seg-min-probesets-CN-0	$\geq 25$

**Figura 45.** Valores mínimos de qualidade para MAPD e *WavinessSD* (1) e Valores mínimos para os filtros de chamadas de CNV (2).

Para as métricas *WavinessSD* e *MAPD*, todas as 841 amostras obtiveram valores abaixo do limite estabelecido de 0,1 para *WavinessSD* e de 0,35 para *MAPD*, assim, todas as amostras seguiram para as próximas etapas do pipeline.

- **Chamadas de CNV por cromossomo**

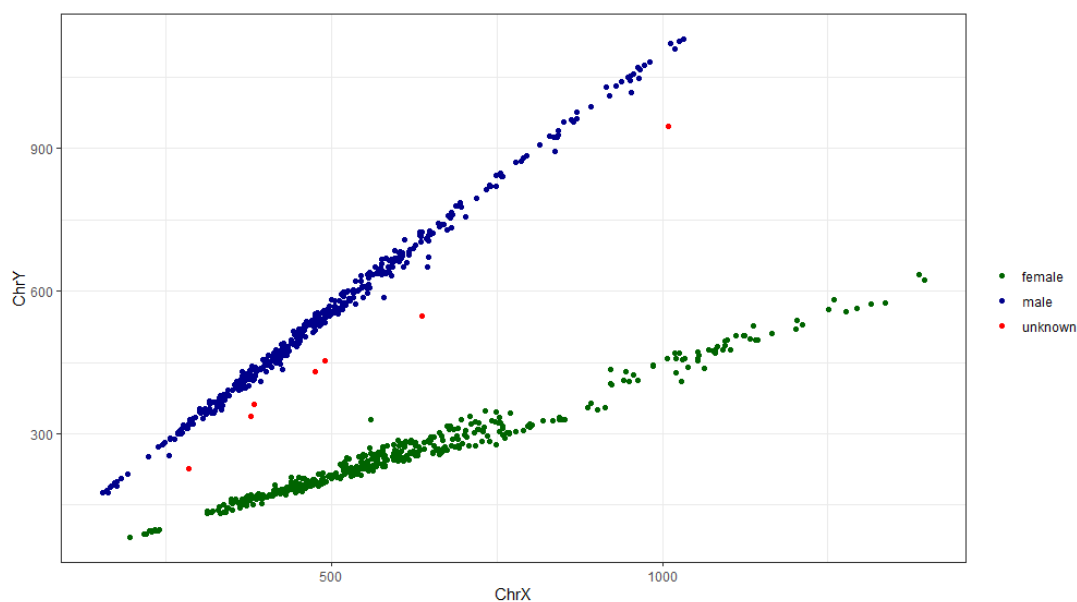
Foi avaliada a distribuição de CNVs ao longo dos cromossomos. O cromossomo Y possui apenas alguns marcadores para a determinação do gênero das amostras e não possui densidade suficiente para a chamada de CNV e LOH, durante o protocolo de chamada de CNV essas regiões são automaticamente ignoradas. O cromossomo 1, o maior cromossomo humano, concentrou a maior parte das chamadas. Contudo, alguns cromossomos menores como 11, 16, 17 e 22, também tiveram um alto número de chamadas. No total foram identificadas 37.217 chamadas. Estudos futuros podem investigar em detalhe o conteúdo das regiões genômicas para verificar se são locais com alta taxa de polimorfismo.



**Figura 46.** Número de CNVs detectadas por cromossomo.

- **Comparação do Gênero das Amostras**

Com base nos dados genômicos avaliados em marcadores nas regiões dos cromossomos sexuais é possível categorizar o gênero dos indivíduos via algoritmo genético. Como uma etapa adicional de verificação de qualidade, foi avaliada a detecção do gênero das amostras computadas pelo algoritmo, o qual compara a razão das expressões dos marcadores SNP do cromossomo X (chrX) e do cromossomo Y (chrY).



**Figura 47.** Comparação dos valores computados do cromossomo X (chrX) e do cromossomo Y (chrY) nas 841 amostras do estudo.

**Tabela 10.** Distribuição dos 841 indivíduos do projeto ISA-Nutrição nas categorias de gênero autodeclaradas e identificadas pelo algoritmo.

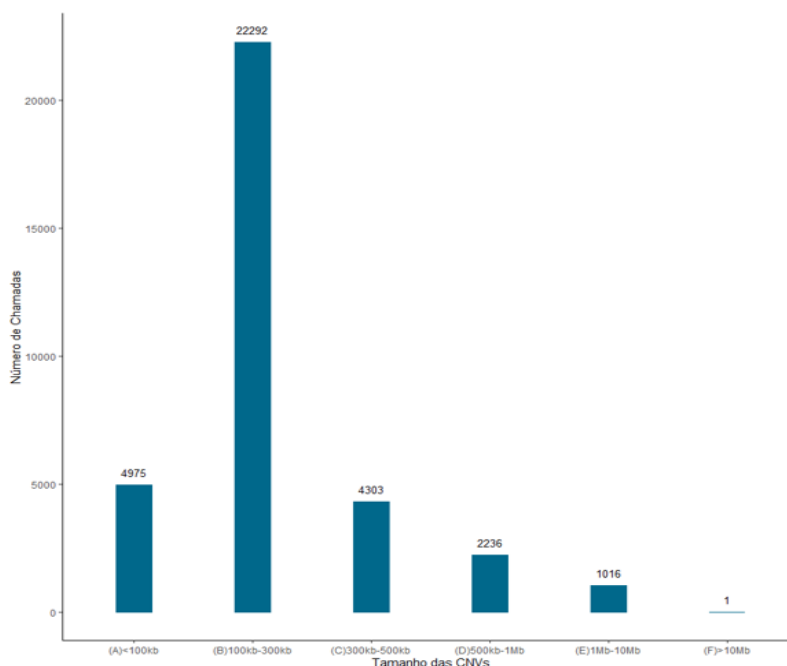
Gênero autodeclarado	Feminino	Masculino	Desconhecido	Total
Feminino	414	4	0	418
Masculino	3	413	7	423
<b>Total</b>	417	417	7	<b>841</b>

Das 841 amostras, 7 foram classificadas como desconhecido (*Unknown*), em vermelho, na Figura 47. Nesses casos, não foi possível classificar adequadamente o gênero dessas amostras. Comparando com o gênero autodeclarado, todas as amostras são do sexo masculino e, pela dispersão dos dados no gráfico, é possível verificar que essas amostras ficaram mais próximas ao limite estabelecido pelo algoritmo para os valores mínimos do cromossomo Y. Também 3 amostras tiveram divergência entre o gênero declarado e identificado pelo algoritmo, para estas amostras foi realizado um levantamento em outros bancos dados do projeto e realizado um cruzamento de dados bioquímicos como testes hormonais e a razão da divergência foi equívocos na categorização do gênero referido, assim os dados foram corrigidos e estão de acordo com os dados genéticos.

- **Tamanho das CNVs**



Foi adotado ao final do protocolo de execução o tamanho mínimo de 25kb para deleções com número de cópias (CN) igual 0, e 50kb para a detecção de deleções com CN = 1 e duplicações (Figura 49). O tamanho das CNVs variou de 49,3kb (CN=0) a 33,5Mb (CN=3) nas amostras de estudo, totalizando 37.217 CNVs para todas as 841 amostras. Para contabilizar o tamanho mais frequente das chamadas, estas foram separadas em 6 categorias de acordo com o tamanho: (A) menores que 100kb, (B) entre 100 e 300kb, (C) entre 300 e 500kb, (D) entre 500kb e 1Mb, (E) entre 1Mb e 10Mb e (F) maiores que 10Mb, conforme Figura 48.



**Figura 48.** Categorias de tamanho das CNVs e os valores de chamadas detectadas durante o processamento.

**Tabela 11.** Estatísticas dos tamanhos (em pares de base – bp) das CNVs por cromossomo (chr).

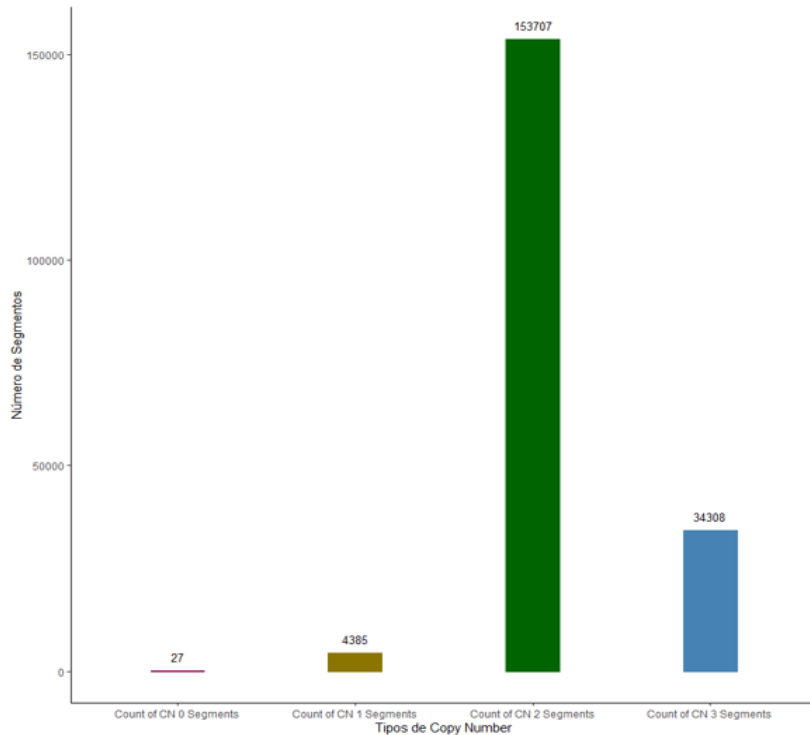
	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
<b>Min.</b>	50.031	50.822	50.181	50.383	51.865	52.039	50.007	50.291
<b>1st Qu.</b>	125.412	127.612	127.857	118.452	123.802	124.228	116.427	120.966
<b>Median</b>	182.414	182.516	184.161	171.797	176.642	182.613	176.549	178.072
<b>Mean</b>	282.913	294.279	326.730	275.863	289.915	298.840	294.965	282.374
<b>3rd Qu.</b>	313.499	296.130	339.153	293.705	302.648	296.268	295.663	303.598
<b>Max.</b>	4.080.382	4.754.442	4.820.484	5.806.174	5.833.755	6.126.949	33.492.830	7.286.792
<b>n</b>	3482	1667	1364	1579	1104	1163	1565	1460
	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16
<b>Min.</b>	50.456	50.541	51.269	51.795	50.794	50.024	51.818	49.286
<b>1st Qu.</b>	119.197	121.928	122.056	125.271	120.843	121.881	124.583	121.758
<b>Median</b>	177.319	180.849	182.993	187.674	178.612	181.898	183.962	186.589

<b>Mean</b>	275.504	285.362	294.947	324.410	367.985	304.468	340.233	286.983
<b>3rd Qu.</b>	293.540	308.131	314.732	330.323	327.437	318.440	322.412	306.844
<b>Max.</b>	9.368.077	2.908.736	3.003.527	11.301.458	13.511.047	15.434.952	18.001.765	20.406.714
<b>n</b>	2082	1698	2588	1290	440	1260	891	2443
	<b>chr17</b>	<b>chr18</b>	<b>chr19</b>	<b>chr20</b>	<b>chr21</b>	<b>chr22</b>	<b>chrX</b>	
<b>Min.</b>	50.480	52.132	50.868	50.232	51.311	50.076	50.420	
<b>1st Qu.</b>	125.086	126.641	118.989	121.233	118.354	121.342	119.030	
<b>Median</b>	191.440	187.986	173.174	183.711	173.258	178.971	169.376	
<b>Mean</b>	321.465	764.126	397.799	343.863	357.240	309.699	337.738	
<b>3rd Qu.</b>	312.828	335.020	264.464	327.411	271.785	273.688	264.454	
<b>Max.</b>	26.364.131	46.373.167	55.776.342	61.368.662	62.400.380	63.006.643	91.016.955	
<b>n</b>	2474	509	1829	1271	665	1999	2394	

As CNVs entre 100kb e 300kb, foram as mais frequentes no estudo. De acordo com estudos prévios (Ning et al., 2022), CNVs pequenas são mais desafiadoras para os métodos e algoritmos atuais, onde ocorrem uma alta taxa de chamadas inespecíficas, muito provavelmente derivadas de ruídos na técnica. De outro lado, as CNVs maiores podem ter um desafio adicional para o desenho experimental, pois exigem uma alta densidade de marcadores ao longo de toda a extensão do evento, o que tem melhorado substancialmente com o avanço das novas tecnologias e com o uso de sondas em regiões contempladas com genes importantes, contudo ainda é um processo custoso (Haraksingh et al., 2017).

- **Tipos de CNV**

O tipo de CNVs mais frequente foi as duplicações de segmentos cromossômicos (Figura 50). Com o ajuste dos dados com uma referência personalizada de amostras do próprio experimento, espera-se diminuir consideravelmente os ruídos derivados de técnica, como do processamento laboratorial e do desempenho das diferentes placas de laboratoriais, que deriva um alto número de chamadas, inclusive de deleções. Alguns estudos prévios também mostraram valores mais altos de duplicações (Sudmant et al., 2015). Como relatado por Brewer et al. (1998), deleções exibem pressões seletivas mais fortes e seus fenótipos costumam ser mais graves em comparação com as duplicações. Por outro lado, as duplicações podem por vezes serem polimorfismos e são utilizadas para estratificação de populações, como adotado na pesquisa de Sudmant et al. (2015), que explorou a diversidade e assinaturas seletivas de deleções e duplicações do tipo CNV em indivíduos de diversas populações humanas, e notou a importância das duplicações para as análises.



**Figura 49.** Número de Cópias por Segmento Cromossômico. O algoritmo separou as deleções nas classes CN=0, CN=1, cópias normais CN=2 e CN=3 para as duplicações.

**Tabela 12.** Frequência de cada tipo de CN (Número de Cópia) por cromossomo (chr).

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
<b>CN=0</b>	4	0	4	0	0	0	0	0
<b>CN=1</b>	170	53	39	38	24	37	17	49
<b>CN=3</b>	3308	1614	1321	1541	1080	1126	1548	1411
	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16
<b>CN=0</b>	0	0	0	0	0	0	0	0
<b>CN=1</b>	87	28	22	19	8	71	16	212
<b>CN=3</b>	1995	1670	2566	1271	432	1189	875	2231
	chr17	chr18	chr19	chr20	chr21	chr22	chrX	
<b>CN=0</b>	0	0	10	0	0	0	9	
<b>CN=1</b>	0	0	0	0	0	0	94	
<b>CN=2</b>	19	14	69	34	8	41	1713	
<b>CN=3</b>	2455	495	1750	1237	657	1958	578	

- **Porcentagem de LOH nos cromossomos autossômicos**

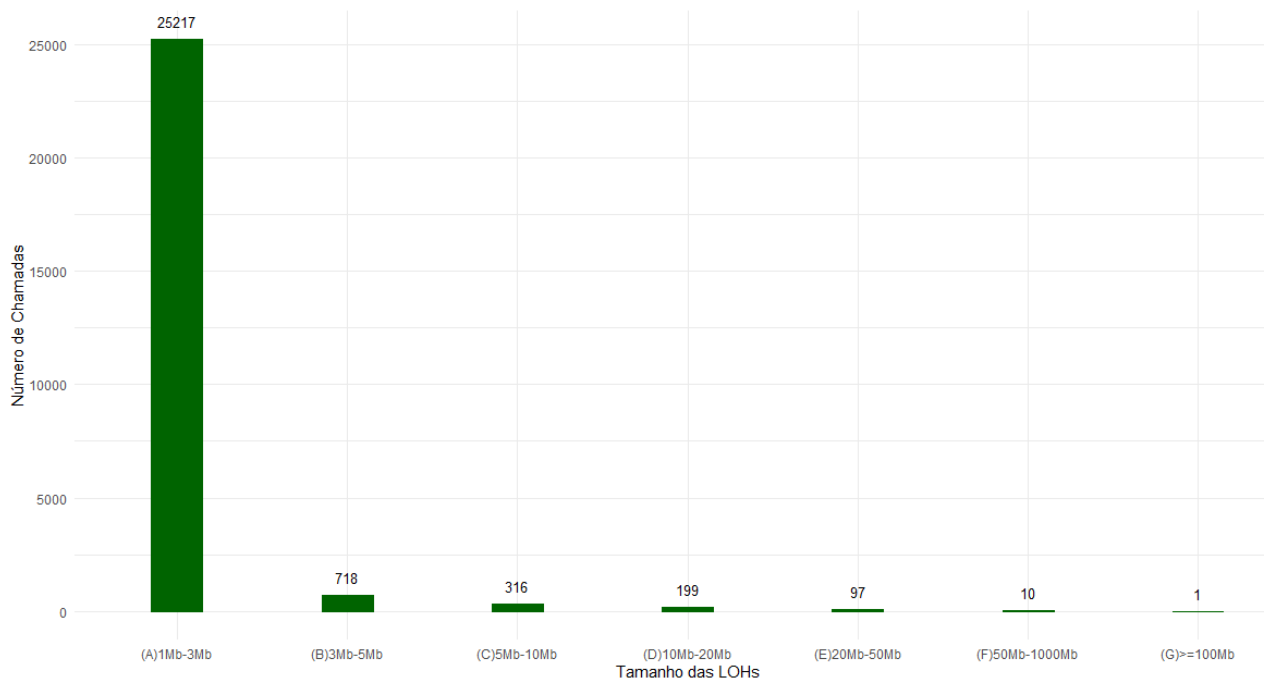
Como saída do protocolo de chamadas de CNV/LOH também é possível estimar a porcentagem de LOHs nos cromossomos autossômicos. Este valor é importante para verificar as amostras com um número maior de regiões com perda de heterozigidade, tanto derivadas por deleções como por aumento da homozigose. As amostras tiveram uma média de 1,565%, variando de 0,285% a 15,73% na porcentagem de LOHs (Tabela 10). Para os valores mais altos, estudos adicionais podem investigar um possível parentesco, ou a qual população ancestral estes indivíduos pertencem.

**Tabela 13.** Estatísticas resumo da % de LOH nos cromossomos autossômicos.

<b>Métrica</b>	<b>Valor (%)</b>
Mínimo	0,285
1 Quartil	1,222
Mediana	1,565
Média	1,989
3 Quartil	2,081
Máximo	15,739

- **Tamanhos das LOHs**

Nas chamadas de LOH, foi aplicado um filtro de mínimo de 25 marcadores às LOHs detectadas. O tamanho das LOHs variou de 1Mb a 108Mb nas amostras de estudo, totalizando 26.558 LOHs para todas as 841 amostras. Para contabilizar o tamanho mais frequente das chamadas, estas foram separadas em 7 categorias de acordo com o tamanho: (A) entre 1Mb e 3Mb, (B) entre 3Mb e 5Mb, (C) entre 5Mb e 10Mb, (D) entre 10Mb e 20Mb, (E) entre 20Mb e 50Mb e (F) entre 50Mb e 100Mb, e (G) maiores que 100Mb, conforme Figura 52.



**Figura 50.** Categorias de tamanho das LOHs e os valores de chamadas detectadas durante o processamento.

**Tabela 14.** Estatísticas dos tamanhos (pares de base – bp) das LOHs por cromossomo (chr).

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
<b>Min.</b>	1.000.345	1.000.039	1.000.069	1.000.034	1.000.262	1.000.038	1.000.949	1.000.247
<b>1st Qu.</b>	1.111.034	1.115.128	1.107.733	1.097.737	1.111.200	1.108.369	1.121.687	1.146.634
<b>Median</b>	1.285.547	1.303.116	1.295.646	1.249.618	1.281.634	1.263.329	1.298.064	1.365.294
<b>Mean</b>	1.716.206	1.693.321	1.751.310	1.791.198	1.702.782	1.682.040	1.611.564	1.797.736
<b>3rd Qu.</b>	1.639.222	1.616.593	1.672.718	1.536.389	1.629.084	1.574.162	1.622.308	1.743.092
<b>Max.</b>	76.085.251	69.177.315	96.801.023	108.789.077	35.397.200	51.600.362	29.828.910	43.224.023
<b>n</b>	2033	2718	2374	1846	1975	1885	1697	1663
	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16
<b>Min.</b>	1.000.447	1.000.357	1.000.166	1.000.360	1.000.698	1.000.145	1.000.796	1.000.159
<b>1st Qu.</b>	1.075.416	1.126.102	1.157.980	1.133.114	1.080.961	1.117.116	1.152.917	1.157.685
<b>Median</b>	1.206.279	1.320.170	1.413.033	1.339.124	1.225.229	1.278.546	1.309.754	1.529.218
<b>Mean</b>	1.848.674	1.698.455	2.143.634	1.685.915	1.691.634	1.693.596	1.793.076	2.176.885
<b>3rd Qu.</b>	1.469.665	1.680.408	2.193.064	1.646.427	1.487.052	1.533.817	1.725.478	2.109.616
<b>Max.</b>	27.616.072	48.634.304	62.710.126	31.163.195	37.132.979	43.762.092	70.553.332	35.026.437
<b>n</b>	625	1308	1483	1368	750	785	910	642
	chr17	chr18	chr19	chr20	chr21	chr22		
<b>Min.</b>	1.000.367	1.000.530	1.001.759	1.000.052	1.007.459	1.000.313		
<b>1st Qu.</b>	1.113.059	1.103.168	1.136.785	1.099.059	1.109.946	1.111.802		
<b>Median</b>	1.284.314	1.240.565	1.329.320	1.259.344	1.239.912	1.273.964		
<b>Mean</b>	1.717.037	1.941.456	1.723.399	1.625.623	1.675.243	1.497.471		

<b>3rd Qu.</b>	1.682.059	1.458.861	1.676.193	1.621.912	1.582.648	1.560.590
<b>Max.</b>	40.405.240	48.986.451	23.932.103	18.011.069	15.544.193	10.070.663
<b>n</b>	750	408	330	538	98	371

Em comparação com o tamanho das CNVs, as LOHs apresentam tamanhos bem maiores, chegando em 108Mb, enquanto a maior CNV tem 33,5Mb. Trabalhos anteriores que estudaram as regiões de perda de heterozigidade/aumento de homozigose, já sugerem que essas regiões podem atender uma escala genômica grande (Pemberton et al., 2012).

- **Sobreposição das Chamadas de CNVs e LOH**

Com a execução do protocolo de sobreposição das chamadas, foram construídas três matrizes, uma contempla a sobreposição de todas as regiões de deleções de CNVs, outra com a sobreposição de todas as regiões de duplicações de CNVs e uma última com a sobreposição de LOHs. As colunas contemplam: cromossomo, posição inicial, posição final, número de sobreposições, lista de amostras do ISA-Nutrição em sobreposição e o status de sobreposição de cada amostra (0 = sem sobreposição e 1 = sobreposição de pelo menos 1bp). Nas linhas, estão todas as regiões genômicas com as CNVs/LOHs com a sobreposição mínima. Com estas informações é possível encontrar a frequência de cada região para todas as amostras do projeto, bem como construir para a amostra ISA-Nutrição a codificação de sobreposição presente ou ausente para todas as regiões identificadas de CNVs (deleção e/ou duplicação) e LOHs. Estes arquivos foram construídos e estão disponíveis para os pesquisadores do ISA.

#### **4.2.3. Avaliação de possíveis fenótipos associados às regiões genômicas dos achados mais frequentes**

- **Identificação das regiões genômicas com a maior frequência de CNVs e LOH**

A região genômica mais deletada nos bancos de dados foi a chr14:20295728-20326905, em 44 dos 841 indivíduos. É uma região próxima ao centrômero do cromossomo 14, e o início do trecho contempla um trecho do último exon do gene OR4N2 e a região 5'UTR. De acordo com a literatura, esse gene faz parte de uma família de proteínas receptoras olfativas que interagem com as moléculas odoríferas do nariz (Malnic et al., 2004).

A região genômica mais duplicada nos bancos de dados foi a chr1:3428696-3448835, presente em 197 dos 841 participantes. A região está em um trecho distal do cromossomo 1, e pega um trecho do gene MEGF6, que foi associado à atividade de ligação do íon cálcio na região

extracelular. Este gene tem ampla expressão na pele e pulmão e quando ocorre a perda de função associada como um fator de predisposição à osteoporose (Teerlink et al., 2021).

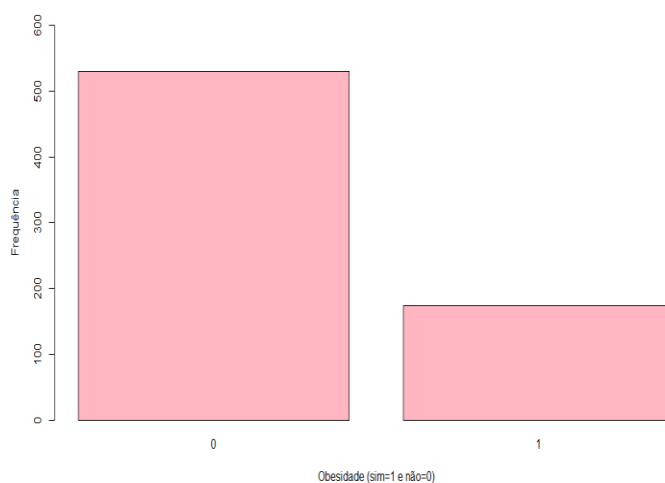
A região genômica de LOH mais frequente nos bancos de dados foi a chr11:49258171-50057150, presente em 250 de 841 indivíduos. A região está em um trecho próximo ao centrômero do cromossomo 11 e contempla um trecho do gene OR4C13. Este gene, também pertence à família de proteínas receptoras olfativas, presentes em vários cromossomos humanos (Malnic et al., 2004).

#### 4.2.4. Associação de % LOHs nos cromossomos autossômicos com o fenótipo de obesidade

Como uma ilustração de uma primeira análise de associação com os dados obtidos, utilizamos o fenótipo de obesidade, atualmente muito estudado por pesquisadores do projeto ISA-Nutrição para determinar os componentes genéticos envolvidos, e comparamos com a porcentagem de LOHs nos cromossomos autossômicos (%LOH).

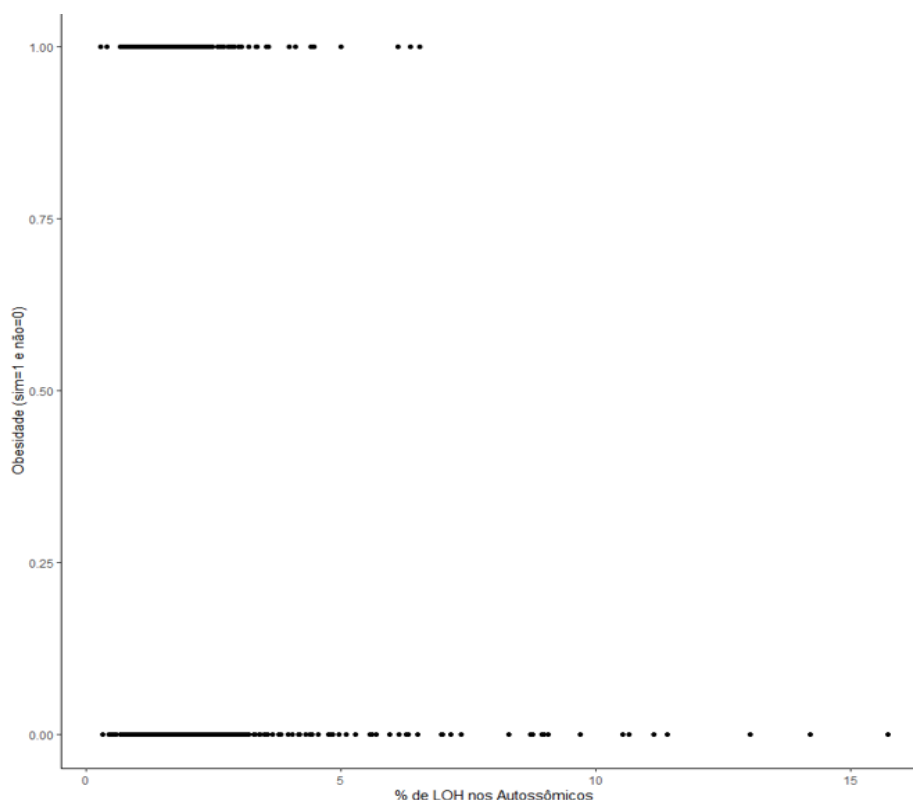
Para categorizar a obesidade, foi utilizado como referência o IMC (Índice de Massa Corpórea,  $IMC = \text{Peso}/\text{Altura}^2$ ) dos indivíduos do estudo, um parâmetro recomendado pela Organização Mundial da Saúde (OMS). Segundo a OMS, valores de IMC acima de  $30\text{Kg}/\text{m}^2$ , já categoriza o indivíduo como obeso. Com esta informação obtida, foi criada uma variável dicotomizada obesidade e todos os indivíduos com o valor de IMC abaixo de  $30\text{Kg}/\text{m}^2$ , foram categorizados como não obesos (não=0) e os indivíduos com valores acima foram classificados como obesos (sim=1).

Para avaliar possíveis variações entre a % de LOH nos cromossomos autossômicos e o índice de obesidade, foi ajustado um modelo de regressão logística adotando a obesidade dicotomizada como variável dependente. Considerando, que alguns participantes do projeto são relacionados, esta avaliação considerou apenas o subconjunto de 707 indivíduos como base.



**Figura 51.** Frequência do Fenótipo Obesidade nas 707 amostras do Projeto-ISA

Para ajustar o modelo de regressão, foram consideradas na análise, além da %de LOH, as seguintes variáveis: sexo dicotomizado (0 = masculino e 1 = feminino), idade ajustada (idade – a média de idade do conjunto) e interação de sexo e idade (sexo \* idade) e os dois primeiros componentes principais de ancestralidade (CP1 e CP2).



**Figura 52.** Gráfico de pontos com a categoria de obesidade e o valor % de LOH nos cromossomos autossômicos.

**Tabela 15.** Valores dos coeficientes da Regressão Logística para Obesidade.

	Estimate	Std.Error	zvalue	Pr(> z )	
<b>(Intercept)</b>	-2,1305	0,27404	-7,774	7,58E-15	***
<b>CP1</b>	-0,0282	0,00512	-5,502	3,75E-08	***
<b>CP2</b>	-14,508	11,766	-1,233	0,21757	
<b>sexo</b>	0,71099	0,19205	3,702	0,00021	***
<b>idade</b>	0,00959	0,00665	1,441	0,14954	
<b>LOH</b>	-0,1371	0,07788	-1,76	0,07844	.
<b>Fator(sexo)*idade</b>	0,01122	0,00897	1,251	0,21077	



Ao avaliar os resultados descritivos (Figura 52), parece existir uma tendência negativa da %LOH nos cromossomos autossômicos e obesidade, onde os valores mais altos de %LOH estão apenas nos indivíduos não obesos. O ajuste do modelo de regressão logística confirma este encontro, em que, a cada aumento de 1% na porcentagem de LOH nos autossomos há uma redução esperada na chance de obesidade de 13% ( $\exp(-0,1371) = 0,87$ ), mantidas fixas as demais variáveis do modelo. Este resultado é significativo estatisticamente a um nível  $\alpha=8\%$ .

As variáveis sexo e CP1, apresentaram valores altamente significantes (valor- $p < 1\%$ ). De acordo com a Figura 33, grupos com maior ancestralidade africana têm valores mais altos do coeficiente CP1 e, portanto, é esperado terem menor chance de obesidade comparados aos grupos com maior ancestralidade europeia, bem como Ameríndia e Asiática. Quanto ao sexo, é esperado que o grupo feminino tenha maior chance de obesidade que o masculino, mantidas fixas as demais variáveis do modelo. Interessantemente, no modelo que inclui somente a %LOH na predição da obesidade o nível descritivo do efeito de LOH foi igual a valor- $p=0,2406$  ( $\beta=-0,0687$ ), indicando o impacto da inclusão das demais variáveis (principalmente, CP1 e sexo) no ganho em precisão da análise.

## 5. Considerações Finais

A motivação deste trabalho está diretamente vinculada ao processamento e análises dos dados genômicos da plataforma *Axiom™ 2.0 Precision Medicine Research Array* do Projeto ISA-Nutrição. A principal contribuição foi processar os dados de uma amostra de brasileiros do projeto ISA-Nutrição para uma caracterização da ocorrência das variantes CNVs e LOH na população brasileira. Um dos produtos deste esforço foi construir e disponibilizar os bancos de dados de genótipos de marcadores SNPs, de regiões de deleções de CNVs, de regiões de duplicação de CNVs e de LOHs, e da análise de ancestralidade global, cujas informações poderão ser usadas em estudos futuros, como a associação destes achados com fenótipos de interesse.

O processamento envolveu alguns desafios, pela estrutura dos dados sob análise. Por exemplo, os dados ISA-Nutrição foram coletados por domicílio e apresentam um grau de parentesco entre alguns participantes. Para acomodar esse padrão dos dados na análise de ancestralidade via Análise de Componentes Principais, foi necessário particionar os dados em amostra de indivíduos independentes e com algum grau de parentesco. Para finalidade deste trabalho, isto foi feito de forma factual categorizando os indivíduos por domicílio e escolhendo um representante a partir do relacionamento com o chefe de domicílio. Para análises futuras, e de forma completamente analítica, a recomendação é construir a matriz de parentesco genético (Wang et al., 2017) entre os indivíduos na amostra e estabelecer um ponto de corte para declaração de dependências.

Para a análise de CNVs e LOHs, dependendo da aplicação, poderá ser necessária etapas extras de filtragem de qualidade, os bancos foram criados de modo mais sensível com cortes mais permissivos de HWE e MAF, permitindo aos pesquisadores flexibilidade das escolhas. Finalmente, neste trabalho fizemos uma tentativa inicial, ilustrativa de associar LOHs com um fenótipo específico, a obesidade. A natureza deste tipo de análise envolve uma estreita colaboração de pesquisa entre o pesquisador da área factual e o bioinformata. Os resultados do presente trabalho correspondem à execução de um primeiro passo neste tipo de pesquisa.

Todos os protocolos de processamento e comandos utilizados durante a execução deste trabalho estão disponíveis no repositório no github em: [https://github.com/camila-alves/scripts\\_ISA](https://github.com/camila-alves/scripts_ISA) e indicados no Apêndice 1.

## 6. Referências

- Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. In *Nature Reviews Genetics* (Vol. 12, Issue 5, pp. 363–376). <https://doi.org/10.1038/nrg2958>
- Alves, M. C. G. P., Escuder, M. M. L., Goldbaum, M., Barros, M. B. de A., Fisberg, R. M., & Cesar, C. L. G. (2018). Sampling plan in health surveys, city of São Paulo, Brazil, 2015. *Revista de Saude Publica*, *52*, 1–11. <https://doi.org/10.11606/S1518-8787.2018052000471>
- Attiyeh, E. F., Diskin, S. J., Attiyeh, M. A., Mossé, Y. P., Hou, C., Jackson, E. M., Kim, C., Glessner, J., Hakonarson, H., Biegel, J. A., & Maris, J. M. (2009). Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Research*, *19*(2), 276–283. <https://doi.org/10.1101/gr.075671.107>
- Barrett, J. C., & Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics*, *38*(6), 659–662. <https://doi.org/10.1038/ng1801>
- Brewer, C., Holloway, S., Zawalnyski, P., Schinzel, A., & Fitzpatrick, D. (1998). A Chromosomal Deletion Map of Human Malformations. In *Am. J. Hum. Genet* (Vol. 63).
- Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, *8*(12). <https://doi.org/10.1371/journal.pcbi.1002822>
- Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 1–16. <https://doi.org/10.1186/s13742-015-0047-8>
- Cordell, H. J., & Clayton, D. G. (2005). Genetic association studies. *Lancet*, *366*(9491), 1121–1131. [https://doi.org/10.1016/S0140-6736\(05\)67424-7](https://doi.org/10.1016/S0140-6736(05)67424-7)
- Cortés-Ciriano, I., Lee, J. J. K., Xi, R., Jain, D., Jung, Y. L., Yang, L., Gordenin, D., Klimczak, L. J., Zhang, C. Z., Pellman, D. S., Akdemir, K. C., Alvarez, E. G., Baez-Ortega, A., Beroukhim, R., Boutros, P. C., Bowtell, D. D. L., Brors, B., Burns, K. H., Campbell, P. J., ... Zhang, C. Z. (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, *52*(3), 331–341. <https://doi.org/10.1038/s41588-019-0576-7>
- D. Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, *3*(25), 731. <https://doi.org/10.21105/joss.00731>
- de Araújo Lima, L., & Wang, K. (2017). PennCNV in whole-genome sequencing data. *BMC Bioinformatics*, *18*(Suppl 11). <https://doi.org/10.1186/s12859-017-1802-x>
- de Onis, M., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., & Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. *Bulletin of the World Health Organization*, *85*(9), 660–667. <https://doi.org/10.2471/BLT.07.043497>
- Deryusheva, I. v, Tsyganov, M. M., Garbukov, E. Y., Ibragimova, M. K., Kzhyshkovska, J. G., Slonimskaya, E. M., Cherdyntseva, N. v, & Litviakov, N. v. (2017). GENOME-WIDE ASSOCIATION STUDY OF LOSS OF HETEROZYGOSITY AND METASTASIS-FREE

SURVIVAL IN BREAST CANCER PATIENTS. In *Experimental Oncology* (Vol. 39).  
<http://omim.org/>

- Dóra, C., Pereira, A., Pacheco, A. G., Santos, R. V., Mill, G., Molina, M., Giatti, L., Almeida, M., Aquino, L., Bensenor, I., & Lotufo, P. A. (2019). Context-dependence of race self-classification : Results from a highly mixed and unequal middle-income country. *PLOS ONE*, 1–17.
- Eddy, S. R. (2004). P R I M E R What is a hidden Markov model? Statistical models called hidden Markov models are a recurring theme in computational biology. What are hidden Markov models, and why are they so useful for so many different problems? 5 I E Start End Figure 1 A toy HMM for 5' splice site recognition. See text for explanation. In *\_computational BIOLOGY* (Vol. 22). <http://www.nature.com/naturebiotechnology>
- Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics*, 14(5), 305–314. <https://doi.org/10.1093/bfgp/elv014>
- Fisberg, R. M., Sales, C. H., De Mello Fontanelli, M., Pereira, J. L., Alves, M. C. G. P., Escuder, M. M. L., César, C. L. G., & Goldbaum, M. (2018). 2015 health survey of São Paulo with focus in nutrition: Rationale, design, and procedures. *Nutrients*, 10(2), 1–13. <https://doi.org/10.3390/nu10020169>
- Ganapathiraju, M. K., Subramanian, S., Chaparala, S., & Karunakaran, K. B. (2020). A reference catalog of DNA palindromes in the human genome and their variations in 1000 Genomes. *Human Genome Variation*, 7(1). <https://doi.org/10.1038/s41439-020-00127-5>
- Giolo, S. R., Soler, J. M. P., Greenway, S. C., Almeida, M. A. A., De Andrade, M., Seidman, J. G., Seidman, C. E., Krieger, J. E., & Pereira, A. C. (2012). Brazilian urban population genetic structure reveals a high degree of admixture. *European Journal of Human Genetics*, 20(1), 111–116. <https://doi.org/10.1038/ejhg.2011.144>
- Haraksingh, R. R., Abyzov, A., & Urban, A. E. (2017). Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. *BMC Genomics*, 18(1). <https://doi.org/10.1186/s12864-017-3658-x>
- Khoury, M. J., & Dotson, W. D. (2021). From genes to public health: are we ready for DNA-based population screening? *Genetics in Medicine*, 23(6), 996–998. <https://doi.org/10.1038/s41436-021-01141-w>
- Kloosterman, W. P., Guryev, V., van Roosmalen, M., Duran, K. J., de Bruijn, E., Bakker, S. C. M., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M., & Cuppen, E. (2011). Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human Molecular Genetics*, 20(10), 1916–1924. <https://doi.org/10.1093/hmg/ddr073>
- Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, 131(7), 1235–1247. <https://doi.org/10.1016/j.cell.2007.11.037>
- Lee, K. W., Woon, P. S., Teo, Y. Y., & Sim, K. (2012). Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: What have we learnt? *Neuroscience and Biobehavioral Reviews*, 36(1), 556–571. <https://doi.org/10.1016/j.neubiorev.2011.09.001>

- Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Vance Morgan, T., Kane, J. M., Kucherlapati, R., & Malhotra, A. K. (2007). *Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia*. [www.pnas.org/cgi/content/full/](http://www.pnas.org/cgi/content/full/)
- Linck, E., & Battey, C. J. (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), 639–647. <https://doi.org/10.1111/1755-0998.12995>
- Lupski, J. R. (2016). *Structural Variation Mutagenesis of the Human Genome: Impact on Disease and Evolution*. *56*(5), 419–436. <https://doi.org/10.1002/em.21943>. Structural
- Malnic, B., Godfrey, P. A., Buck, L. B., & Howard, S. J. (2004). *The human olfactory receptor gene family*. [www.pnas.org/cgi/doi/10.1073/pnas.0307882100](http://www.pnas.org/cgi/doi/10.1073/pnas.0307882100)
- Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), 1–10. <https://doi.org/10.1002/mpr.1608>
- Mersha, T. B., & Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: Its potential impact on understanding health disparities. *Human Genomics*, *9*(1), 1–15. <https://doi.org/10.1186/s40246-014-0023-x>
- Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J. A., Albrechtsen, A., Malaspina, A. S., Sikora, M., Reuther, J. D., Irish, J. D., Malhi, R. S., Orlando, L., Song, Y. S., Nielsen, R., Meltzer, D. J., & Willerslev, E. (2018). Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, *553*(7687), 203–207. <https://doi.org/10.1038/nature25173>
- Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., Barnes, M. R., Barnett, A. H., Bates, C., Bellary, S., Bockett, N. A., Giorda, K., Griffiths, C. J., Hemingway, H., Jia, Z., Kelly, M. A., Khawaja, H. A., Lek, M., McCarthy, S., McEachan, R., ... Van Heel, D. A. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science*, *352*(6284), 474–477. <https://doi.org/10.1126/science.aac8624>
- Neher, R. A., & Shraiman, B. I. (2011). Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics*, *188*(4), 975–996. <https://doi.org/10.1534/genetics.111.128876>
- Ning, Y., Czekalski, M., Herrada, S., & Greene, C. (2022). Interpretation challenge of small copy number variations in the imprinting regions. In *Molecular Genetics and Genomic Medicine*. John Wiley and Sons Inc. <https://doi.org/10.1002/mgg3.1961>
- Parks, M. M., Lawrence, C. E., & Raphael, B. J. (2015). Detecting non-allelic homologous recombination from high-throughput sequencing data. *Genome Biology*, *16*(1), 1–19. <https://doi.org/10.1186/s13059-015-0633-1>
- Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, *91*(2), 275–292. <https://doi.org/10.1016/j.ajhg.2012.06.014>
- Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., ... Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations:

- Opportunities, Methods, Pitfalls, and Recommendations. In *Cell* (Vol. 179, Issue 3, pp. 589–603). Cell Press. <https://doi.org/10.1016/j.cell.2019.08.051>
- Pillai, N. E., Okada, Y., Saw, W. Y., Ong, R. T. H., Wang, X., Tantoso, E., Xu, W., Peterson, T. A., Bielawny, T., Ali, M., Tay, K. Y., Poh, W. T., Tan, L. W. L., Koo, S. H., Lim, W. Y., Soong, R., Wenk, M., Raychaudhuri, S., Little, P., ... Teo, Y. Y. (2014). Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Human Molecular Genetics*, 23(16), 4443–4451. <https://doi.org/10.1093/hmg/ddu149>
- Pongpanich, M., Sullivan, P. F., & Tzeng, J. (2010). A quality control algorithm for filtering SNPs in genome-wide association studies. 26(14), 1731–1737. <https://doi.org/10.1093/bioinformatics/btq272>
- Prive, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics*, 34(16), 2781–2787. <https://doi.org/10.1093/bioinformatics/bty185>
- Puig, X., Ginebra, J., & Graffelman, J. (2019). Bayesian model selection for the study of Hardy–Weinberg proportions and homogeneity of gender allele frequencies. *Heredity*, 123(5), 549–564. <https://doi.org/10.1038/s41437-019-0232-0>
- 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>
- Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., ... Hurler, M. E. (2006). Global variation in copy number in the human genome. *Nature*, 444(7118), 444–454. <https://doi.org/10.1038/nature05329>
- Ryland, G. L., Doyle, M. A., Goode, D., Boyle, S. E., Choong, D. Y. H., Rowley, S. M., Li, J., Bowtell, D. D., Tothill, R. W., Campbell, I. G., & Gorringer, K. L. (2015). Loss of heterozygosity: What is it good for? *BMC Medical Genomics*, 8(1), 1–12. <https://doi.org/10.1186/s12920-015-0123-z>
- Santos, H. C., Horimoto, A. V. R., Tarazona-Santos, E., Rodrigues-Soares, F., Barreto, M. L., Horta, B. L., Lima-Costa, M. F., Gouveia, M. H., Machado, M., Silva, T. M., Sanches, J. M., Esteban, N., Magalhaes, W. C. S., Rodrigues, M. R., Kehdy, F. S. G., & Pereira, A. C. (2016). A minimum set of ancestry informative markers for determining admixture proportions in a mixed American population: The Brazilian set. *European Journal of Human Genetics*, 24(5), 725–731. <https://doi.org/10.1038/ejhg.2015.187>
- Secolin, R., Mas-Sandoval, A., Arauna, L. R., Torres, F. R., de Araujo, T. K., Santos, M. L., Rocha, C. S., Carvalho, B. S., Cendes, F., Lopes-Cendes, I., & Comas, D. (2019). Distribution of local ancestry and evidence of adaptation in admixed populations. *Scientific Reports*, 9(1). <https://doi.org/10.1038/s41598-019-50362-2>
- Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B. P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L. B., Posukh, O. L., Sahakyan, H., Watkins, W. S., Yepiskoposyan, L., Abdullah, M. S., Bravi, C. M., Capelli, C., Hervig, T., ... Eichler, E. E. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science*, 349(6253). <https://doi.org/10.1126/science.aab3761>
- Taylor, R. W., & Turnbull, D. M. (2005). Mitochondrial DNA mutations in human disease. *Nature Reviews Genetics*, 6(5), 389–402. <https://doi.org/10.1038/nrg1606>

- Teerlink, C. C., Juryneć, M. J., Hernandez, R., Stevens, J., Hughes, D. C., Brunker, C. P., Rowe, K., Grunwald, D. J., Facelli, J. C., & Cannon-Albright, L. A. (2021). A role for the MEGF6 gene in predisposition to osteoporosis. *Annals of Human Genetics*, *85*(2), 58–72. <https://doi.org/10.1111/ahg.12408>
- The 1000 Genomes Project Consortium. (2015). Processing 1000 Genomes reference data for ancestry estimation. *Nature*, *526*, 68–88. <https://doi.org/https://doi.org/10.1038/nature15393>
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. In *Nature Reviews Methods Primers* (Vol. 1, Issue 1). Springer Nature. <https://doi.org/10.1038/s43586-021-00056-9>
- Wang, B., Sverdlow, S., & Thompson, E. (2017). Efficient estimation of realized kinship from single nucleotide polymorphism genotypes. *Genetics*, *205*(3), 1063–1078. <https://doi.org/10.1534/genetics.116.197004>
- Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., Hakonarson, H., & Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research*, *17*(11), 1665–1674. <https://doi.org/10.1101/gr.6861907>
- Weischenfeldt, J., Symmons, O., Spitz, F., & Korbelt, J. O. (2013). Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics*, *14*(2), 125–138. <https://doi.org/10.1038/nrg3373>
- Yang, H. C., Chen, C. W., Lin, Y. T., & Chu, S. K. (2021). Genetic ancestry plays a central role in population pharmacogenomics. *Communications Biology*, *4*(1). <https://doi.org/10.1038/s42003-021-01681-6>
- Zhang, X., & Sjöblom, T. (2021). Targeting loss of heterozygosity: A novel paradigm for cancer therapy. *Pharmaceuticals*, *14*(1), 1–17. <https://doi.org/10.3390/ph14010057>
- Zheng, X., Gogarten, S. M., Lawrence, M., Stilp, A., Conomos, M. P., Weir, B. S., Laurie, C., & Levine, D. (2017). SeqArray—a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics*, *33*(15), 2251–2257. <https://doi.org/10.1093/bioinformatics/btx145>
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. <https://doi.org/10.1093/bioinformatics/bts606>
- BRASIL. Estatuto da Criança e do Adolescente. Lei 8.069/90. São Paulo, Atlas, 1991.
- BRASIL. Estatuto do Idoso. Lei 10.741/03, Brasília, DF, 2003.

# Apêndice

Todos os protocolos de execução e comandos utilizados durante a execução deste trabalho estão disponíveis no repositório no github em: [https://github.com/camila-alves/scripts\\_ISA](https://github.com/camila-alves/scripts_ISA). Abaixo está o arquivo README.md do repositório com o link dos algoritmos utilizados e suas respectivas versões.

## Scripts\_Mestrado

### Data analysis Cohort - ISA Capital

Samples - Array Axiom Precision Medicine Research Array (PMRA) <https://www.thermofisher.com/order/catalog/product/902981#/902981>

841 individuals

873,177 SNP recommended

---

### Sources:

APT - Affymetrix <https://www.affymetrix.com/support/developer/powertools/changelog/apt-genotype-axiom.html>

Plink 1.9 <https://www.cog-genomics.org/plink/>

Plink 2.0 <https://www.cog-genomics.org/plink/2.0/>

SNPRelate <https://www.bioconductor.org/packages/release/bioc/html/SNPRelate.html>

PennCNV: <https://penncnv.openbioinformatics.org/en/latest/>

---

### Programming Language:

PowerShell

Shell Script

R

Python

---

**Database:** Google Drive - Bancos Processados Genética ISA - Folder