UNIVERSIDADE FEDERAL DE GOIÁS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA
E BIODIVERSIDADE
DOUTORADO EM BIOTECNOLOGIA E BIODIVERSIDADE

DETECÇÃO E ANÁLISE DE MUTAÇÕES *DE NOVO* EM PACIENTES COM EXPOSIÇÃO PARENTAL À RADIAÇÃO IONIZANTE DE CÉSIO-137 A PARTIR DE DADOS DE GENOTIPAGEM DE POLIMORFISMOS DE BASE ÚNICA DE ALTA DENSIDADE

HUGO PEREIRA LEITE FILHO

GOIÂNIA
GOIÁS – BRASIL
©2020

UFG

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE PATOLOGIA TROPICAL E SAÚDE PÚBLICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [X] Tese

2. Nome completo do autor

HUGO PEREIRA LEITE FILHO

3. Título do trabalho

DETECÇÃO E ANÁLISE DE MUTAÇÕES DE NOVO EM PACIENTES COM EXPOSIÇÃO PARENTAL À RADIAÇÃO IONIZANTE DE CÉSIO-137 A PARTIR DE DADOS DE GENOTIPAGEM DE POLIMORFISMOS DE BASE ÚNICA DE ALTA DENSIDADE

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Cláudio Carlos da Silva, Usuário Externo**, em 10/11/2020, às 18:00, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **HUGO PEREIRA LEITE FILHO**, **Discente**, em 11/11/2020, às 20:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1664573** e o código CRC **69F9B348**.

Referência: Processo nº 23070.047161/2020-91 SEI nº 1664573

HUGO PEREIRA LEITE FILHO

DETECÇÃO E ANÁLISE DE MUTAÇÕES *DE NOVO* EM PACIENTES COM EXPOSIÇÃO PARENTAL À RADIAÇÃO IONIZANTE DE CÉSIO-137 A PARTIR DE DADOS DE GENOTIPAGEM DE POLIMORFISMOS DE BASE ÚNICA DE ALTA DENSIDADE

Orientador: Prof. Dr. Cláudio Carlos da Silva
Coorientador: Prof. Dr. Alexandre Rodrigues Caetano

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia e Biodiversidade, para obtenção do título de Doutor.

GOIÂNIA
GOIÁS – BRASIL
©2020

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Pereira Leite Filho, Hugo
Detecção e Análise de Mutações de Novo em Pacientes com Exposição Parental à Radiação Ionizante de Césio-137 a Partir de Dados de Genotipagem de Polimorfismos de Base Única de Alta Densidade [manuscrito] / Hugo Pereira Leite Filho. - 2020.
CXLV, 145 f.: il.

Orientador: Prof. Dr. Cláudio Carlos da Silva; co-orientador Dr. Alexandre Rodrigues Caetano.
Tese (Doutorado) - Universidade Federal de Goiás, Instituto de Patologia Tropical e Saúde Pública (IPTSP), Programa de Pós graduação em Biotecnologia e Biodiversidade, Cidade de Goiás, 2020.
Bibliografia. Anexos. Apêndice.
Inclui gráfico, tabelas, algoritmos, lista de figuras, lista de tabelas.

1. Bioinformática. 2. SNPs. 3. Genotipagem. 4. Desvio Mendeliano. 5. CytoScan 750K. I. Carlos da Silva, Cláudio , orient. II. Título.

CDU 60

UFG

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE PATOLOGIA TROPICAL E SAÚDE PÚBLICA

ATA DE DEFESA DE TESE

ATA DA REUNIÃO DA BANCA EXAMINADORA DA DEFESA DE TESE DE HUGO PEREIRA LEITE FILHO - Aos vinte e oito dias do mês de outubro do ano de 2020 (28/10/2020), às 14:00 horas, reuniram-se os componentes da Banca Examinadora: Profs. Drs. **Cláudio Carlos da Silva** (PUC/GO) (orientador), **Mariana Pires de Campos Telles** (ICB/UFG), **Daniela de Melo e Silva** (ICB/UFG), **Alex Silva da Cruz** (PUC/GO) e **Marc Alexandre Duarte Gigonzac** (PUC/GO) para, sob a presidência do primeiro, e em sessão pública por webconferência, procederem à avaliação da defesa de tese intitulada: "**DETECÇÃO E ANÁLISE DE MUTAÇÕES DE NOVO EM PACIENTES COM EXPOSIÇÃO PARENTAL À RADIAÇÃO IONIZANTE DE CÉSIO-137 A PARTIR DE DADOS DE GENOTIPAGEM DE POLIMORFISMOS DE BASE ÚNICA DE ALTA DENSIDADE**", em nível de **DOUTORADO**, área de concentração em **BIOTECNOLOGIA**, de autoria de **HUGO PEREIRA LEITE FILHO**, discente do PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA E BIODIVERSIDADE, da Universidade Federal de Goiás. A sessão foi aberta pelo orientador do discente, Prof. Dr. CLÁUDIO CARLOS DA SILVA, que fez a apresentação formal dos membros da Banca e orientou o Candidato sobre como utilizar o tempo durante a apresentação de seu trabalho. A palavra a seguir, foi concedida ao autor da tese que, em 30 minutos procedeu à apresentação de seu trabalho. Terminada a apresentação, cada membro da Banca arguiu o Candidato, tendo-se adotado o sistema de diálogo sequencial. Terminada a fase de arguição, procedeu-se à avaliação da defesa. Tendo-se em vista o que consta na Resolução nº. 1181/2013 do Conselho de Ensino, Pesquisa, Extensão e Cultura (CEPEC), que regulamenta o Programa de Pós-Graduação em Biotecnologia e Biodiversidade a Banca, em sessão secreta, expressou seu Julgamento, considerando o candidato Aprovado ou Reprovado:

Banca Examinadora	**Aprovado / Reprovado**
Prof. Dr. Cláudio Carlos da Silva	**Aprovado**
Profa. Dra.Mariana Pires de Campos Telles	**Aprovado**
Profa. Dra. Daniela de Melo e Silva	**Aprovado**
Prof. Dr. Alex Silva da Cruz	**Aprovado**
Prof. Dr. Marc Alexandre Duarte Gigonzac	**Aprovado**

Em face do resultado obtido, a Banca Examinadora considerou o candidato **Habilitado**, (**Habilitado ou não Habilitado**), cumprindo todos os requisitos para fins de obtenção do título de **DOUTOR EM BIOTECNOLOGIA E BIODIVERSIDADE**, na área de concentração em **BIOTECNOLOGIA**, pela Universidade Federal de Goiás. Cumpridas as formalidades de pauta, às _17__h_00__min, a presidência da

mesa encerrou esta sessão de defesa de tese e para constar eu, HELOÍSA DE SOUSA VIEIRA, secretária do Programa de Pós-Graduação em Biotecnologia e Biodiversidade lavrei a presente Ata que depois de lida e aprovada, será assinada pelos membros da Banca Examinadora.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Mariana Pires De Campos Telles, Professor do Magistério Superior**, em 29/10/2020, às 14:38, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Cláudio Carlos da Silva, Usuário Externo**, em 29/10/2020, às 14:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **ALEX SILVA DA CRUZ, Usuário Externo**, em 29/10/2020, às 15:30, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Daniela De Melo E Silva, Professor do Magistério Superior**, em 02/11/2020, às 20:40, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marc Alexandre Duarte Gigonzac, Usuário Externo**, em 05/11/2020, às 11:15, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1625483** e o código CRC **4B3F5DFA**.

Referência: Processo nº 23070.047161/2020-91 SEI nº 1625483

INSTITUIÇÕES PARTICIPANTES

O desenvolvimento das atividades experimentais do presente estudo contou com a colaboração direta das instituições: Núcleo de Pesquisas Replicon da Escola de Ciências Agrárias e Biológicas da Pontifícia Universidade Católica de Goiás (PUC-Goiás), Laboratório de Citogenética Humana e Genética Molecular do Estado de Goiás (LaGene/LACEN/SES-GO). Além disso, o desenvolvimento da tese foi apoiado pelas Universidade Federal de Goiás (UFG) e Universidade de Brasília (UnB) através do Programa de Pós-Graduação em Biotecnologia e Biodiversidade da Rede Centro-Oeste de Pós-Graduação, Pesquisa e Inovação e Embrapa Recursos Genéticos e Biotecnologia.

Por demais, as atividades desenvolvidas tiveram o apoio do Tribunal Regional Federal da 1ª Região.

"O homem é uma animal criador por excelência, condenado a tender conscientemente, para um objetivo e a ocupar-se da arte da engenharia, isto é, abrir para si mesmo um caminho, eterna e incessantemente, para onde quer que seja."

Fiodor Dostoiévski

DEDICATÓRIA

DEDICO Á LORENA, a minha amada, minha fortaleza, companheira de vida e no amor. Contar com seu apoio irrestrito foi o equilíbrio necessário para que esta obra fosse finalizada. Sou eternamente grato por tudo, por nós, pelas conquistas que nos deixaram mais fortes e convencidos que juntos os desafios nos tornam melhores.

DEDICO AOS MEUS PRECIOSOS FILHOS GIOVANNA e LUCAS, minhas dádivas, meus amores incondicionais. Com vocês presentes, me sinto mais evoluído e mais certo que vocês são essenciais em minha vida.

DEDICO À MINHA MÃE MARIA E AO MEU PAI HUGO (*in memoriam*), pela imensa magnitude que vocês representam em minha caminhada, pelo caráter, humildade, proteção e respeito que sempre os acompanharam.

AGRADECIMENTOS

A contribuição de várias pessoas com sugestões, críticas, carinho, apoio, ideias, afeto e muita amizade, foi que proporcionou a realização e finalização deste trabalho. Agradeço:

À minha esposa Lorena Pires Carneiro Leite e aos meus filhos Giovanna Buzolo Leite, Lucas Carneiro Leite, por fazerem parte da vida e se este amor incondicionável e insubstituíveis. Amo e admiro muito vocês!

Aos meus pais Hugo Pereira Leite (*in memorian*) e Maria Pereira Santos, por serem a minha referência como pessoa e ser humano. Tenho muito orgulho de vocês e me sinto abençoada por ser filho de vocês!

Ao Prof. Dr. Aparecido Divino da Cruz, por ser este construtor de sonhos, um idealizador no universo do ensino, da pesquisa, principalmente na formação do pensamento intelectual e crítico. Serei eternamente grato pela confiança que sempre me enxergou. Um grande admirador, desse olhar bondoso e acolhedor. Tenho por ti uma referência a ser seguido, por todo o conjunto que a vossa pessoa contempla (hombridade, respeito e generosidade), sendo extremamente grato pelo nosso vínculo de amizade, carinho e amor construídos durante esta trajetória de vida. Minha mais absoluta gratidão por fazer, querer e insistir nesta grande e outras conquistas.

Ao Prof. Dr. Cláudio Carlos da Silva, por ser este orientador compreensível, dedicado. Por ser este transformador de sonhos em realidade, por essa maestria no campo da sabedoria. Por toda a confiança em mim depositada. Pela nossa amizade solidificada e repleto de respeito. Meu mais intenso agradecimento.

Ao Prof. Dr. Alexandre Rodrigues Caetano, pela prontidão neste desafio da coorientação, pela disponibilidade que sempre se colocou à disposção. Pela colaboração com ideias inovadoras e que deram o rumo para o desenvolvimento da pesquisa. Pelo domínio do conhecimento que tanto soube repassar. Por todos os momentos de imensa ajuda e disposição, meu profunda gratidão.

Às Profª. Drª. Irene Plaza e Profª. Drª. Emília, pelo apoio incondicional, pela paciência e dedicação em todas as etapas desse projeto de vida. Não há palavras para reconhecer a gratidão e carinho que sinto por vocês. Minha total admiração e gratidão por vocês fazerem parte desta conquista.

Ao Profº. Dr. Alex Silva da Cruz, Profª. Drª. Daniele Silva, Profª. Drª. Mariana Telles, Profº. Dr. Marc Gigonzac e a Profª. Drª. Lysa Minasi pelas contribuições, apoio e discussões positivas e construtivas referentes ao estudo, que me fizeram crescer profissionalmente e pessoalmente.

Aos amigos Damiana Míriam da Cruz e Cunha, Cristiano Luiz Ribeiro, Eduardo Rocha Pedrosa por fazer parte desta amizade e conhecimento. Contar com vocês é a mais certeza que conquistas farão parte de nossas vidas. Conviver com vocês é absolutamente agradável.

Aos amigos do Laboratório REPLICON-PUC-GO e LAGENE-SES-GO: Lorraynne Guimarães Oliveira, Ana Júlia Cunha Leite, Marcos Vinícius Milk, Andreia Pires Amancio, Samara Socorro Silva Pereira, Rafael Carneiro, Thaís Cidália Vieira, João Antônio Xavier Manso, Raphael Silva da Cruz, Calebe Bertolino Marins de Campos e Elder Balestra, por partilharem e compartilharem amizade, apoio e aprendizado.

Aos amigos do Tribunal Regional Federal da 1ª Região: Mário Braga, Giuseppe Janino Júnior, Fábio Gaudine, Marco Aurélio Santos, Eduardo Bogoni, Thiago Mota, Alex Pitacci, Gustavo Luís, Gilmar Nonato, Antônio Giovanni, Giscard Stephanou, Marcello Costa, Vírginia Correia, Janderson Santos, Mônica Regina, Marcos Salenko, Alex Alves, Anibal Martins, André Gonçalves, André de Mello, Paulo Cesar Filho, Eduardo Henrique, Regina Pereira, Juliano Vasconcelos, Klayton de Sousa, Sérgio Lisias, José Ferretti, Fabrício Ferreira, Andreia Rodrigues, Sonaira Larissa, Washington Henrique, Kênia Nascimento, Ricardo Andrade, Carla Cristina, Edmar Barboza, Marcelo Pinheiro, M. Juiz Rodrigo de Godoy Mendes e M. Juiz Roberto Carvalho Veloso, pelo apoio incondicional e confiança.

Aos amigos professores colegas da Universidade Estadual de Goiás, Marcelo Ortega, Jeferson Araújo, Pollyana Fanstone, Ly Freitas, Francino Azevedo, Walter Dias Jr., Elton Morais, Guthemberg Rocha e colegas da área administrative do campus Ceres que tanto me apoiaram nesta importante conquista.

Aos meus amigos José Luiz da Cruz, Maria Leonor Furtado, Ana Furtado, Andrey Coelho, Maria Inez Furtado e Marcinha Furtado (*in memorian*), Maria Dalva Furtado (Lalá), Karine Bonfim, Marcelo Castilho, pela torcida e carinho.

A todos os professores do Programa de Pós-Graduação em Biotecnologia e Biodiversidade da Rede Centro-Oeste de Pós-Graduação, Pesquisa e Inovação pela troca de conhecimentos e experiências.

Aos meus queridos e amados irmãos e cunhadas Adriano Pereira Leite, Denise Machado Leite, Anderson Pereira Leite, Nelma Severino por serem quem vocês são e por vivenciarmos nossas vidas com muito amor, admiração, carinho e alegria.

Às minhas sobrinhas e sobrinhos Ítalo Machado Leite, Beatriz Machado Leite, André Leite por me ajudarem a ser uma pessoa melhor.

Ao meu sogro Almir Carneiro, Divina Carneiro, Kátia Eveline (*in* memorian), Hugo Pires Carneiro e Gabriel Carneiro por fazerem parte da minha vida e estarem ao meu lado.

A todos e todas pacientes e familiares que, com a sua humildade, esperança e luta, fizeram aprimorar o meu olhar para o outro e sua condição e a entender que a empatia é a chave para um mundo mais humano. Agradeço também por aceitaram participar da pesquisa e colaborarem com meu aprendizado. Minha respeitosa e eterna gratidão!

Meu cordial e verdadeiro agradecimento a todos vocês.

SUMÁRIO

LISTA DE FIGURAS16
LISTA DE TABELAS, EQUAÇÕES E QUADROS17
RESUMO18
1. INTRODUÇÃO22
2. REVISÃO BIBLIOGRÁFICA24
2.1. Bioinformática24
2.2. Matrizes de Genotipagem de Polimorfismo em Nucleotídeos Único25
2.3. Tecnologia Affymetrix – Genotipagem por SNP26
2.4. Similaridade entre os softwares de mercado: Affymetrix e Illumina29
2.5. Software ChAS Affymetrix31
2.6. Soluções Recursivas33
2.7. Cálculo Proposicional35
2.8. Histórico sobre o acidente radiológico com Césio-13736
2.9. Exposição dos sistemas biológicos à radiação ionizante37
2.10. Efeitos das idades parentais nas mutações39
2.11. Desequilibrio de Ligação41
2.12. PLINK42
2.13. Medidas de Desiquilíbrio de Ligação para Único Par de SNPs43
3 OBJETIVOS44
3.1 Objetivo Geral44
3.2 Objetivos Específicos44
4 METODOLOGIA45
4.1 Delineamento do Estudo45
4.2 Caracterização do Grupo Amostral47
4.3 Considerações Éticas48
4.4 Obtenção dos Dados48
4.5 Análise Cromossômicas por Microarranjos49
4.6 Estimativa dos Desvios Mendelianos49
4.7 Frequência Média de Desvio Mendeliano..... 52
4.8. Prunning de SNP's Baseado no Desequilíbrio de Ligação55
4.9 Análise estatística55
5. RESULTADOS56
6. DISCUSSÃO..... 69
7. CONCLUSÃO..... 74

8. REFERÊNCIA BIBLIOGRÁFICA..... 75
APÊNDICE I..... 86
APÊNDICE II..... 88
APÊNDICE III 92
ANEXO I..... 103
ANEXO II 107
ANEXO III105

LISTA DE FIGURAS

Figura 1. Representando uma visão geral, em forma de fluxograma, da análise do pipeline aplicada ao protocolo CytoScan HD™ 26
Figura 2. Visão geral da representação das etapas construídas pelo algoritmo BRLMM-P-Plus..... 27
Figura 3. Etapas usando a metodologia CMA que está incorporada na plataforma ChAS®..... 30
Figura 4. Descreve o Funcionamento de uma divisão de processamento em um modelo computacional..... 31
Figura 5. Desenha uma série geométrica para *n* recorrências..... 33
Figura. 6. Representa o fluxo lógico para demostrar como as soluções recursivas..... 34
Figura 7. As imagens descrevem o ciclo das células reprodutivas..... 38
Figura 8. *Workflow* do Software ChAS® 46
Figura 9. Fluxo das etapas geradas pelos algoritmos para identificação dos desvios mendelianos por SNP 49
Figura 10. Fluxo das etapas que demonstram as etapas dos algoritmos para os achados de desvios mendelianos para cada trio 52
Figura 11. Fluxo dos procedimentos adotados para a execução dos *pipelines* para detecção de erros mendelianos 53
Figura 12. Gráfico QQ representando os valores dos quartis referente ao número de desvios mendelianos por grupos caso e controle 59
Figura 13. Os representam as médias dos números de desvios mendelianos e médias das taxas de mutação germinativa da progênie de caso e controle, expostos às dosagens baixas de radiação ionizantes do Césio-137 em Goiânia (Brasil) 60
Figura 14. Representam a regressão linear (frequência de mutação vs dose absorvida (Gy)) 61
Figura 15. Efeito potencial da idade parental na concepção em relação ao número de desvios mendeliano 62
Figura 16. Temos a regressão linear do efeito das idades parentais sobre o desvio mendeliano para os grupos caso e controle..... 64
Figura 17. Temos a representação em forma de diagrama de Venn dos números de bases nitrogenadas a partir dos DM 64
Figura 18. Cluster e PCA baseado na distância aos pares da identidade-por-estado (IBS)..... 65
Figura 19. Os gráficos representam a técnica de Tukey para realizar múltiplas comparações usando as variáveis idade e grupo..... 66
Figura 20. Os resultados representam a proporção entre os tipos de substituições: Transição e Transversão..... 67
Figura 21. Mutações de novo em fase baseadas em alelos parentais e derivados distribuídos em classes de substituições de bases para casos e controles..... 68
Figura 22. Indica a proporção de inferências realizadas nos grupos caso e controle, pois as origens dos progenitores em relação às proles eram indentificáveis..... 69

LISTA DE TABELAS, EQUAÇÕES E QUADROS

Tabela 1. Representação das fases dos alelos para os haplótipos sob Equilíbrio de Ligação dos SNPs rs2840528 rs7545940.....41
Tabela 2. Dados gerais dos grupos controle e exposto a respeito do estudo de mutação da linha germinativa em filhos de pessoas acidentalmente expostas a baixas doses absorvidas de radiação ionizante de césio-137 em Goiânia (Brasil).....56
Tabela 3. Resumo dos dados descritivos dos grupos caso e controle para as seis classes de substituição de bases no genoma de crianças concebidas após exposição dos pais a baixas doses de radiação ionizante e seus controles.....66
Quadro 1. Combinações que foram geradas pelo script nas mutações de bases nitrogenadas.....50
Quadro 2. Combinações geradas pelo script nas mutações de genotipagem.....51
Quadro 3. Dados gerais dos grupos caso e controle para as gerações parental e F1 incluídos no estudo da sobre a indução de mutação germinativa na prole de indivíduos expostos acidentalmente a doses baixas de radiação ionizante de césio-137.....55

RESUMO

Em 1987, na cidade de Goiânia, uma série de eventos inesperados resultou em um grave acidente radiológico, gerado pelo Césio-137. Os efeitos mutagênicos da radiação ionizante (RI) pode levar acúmulo de mutações em filhos de pais irradiados. Foi estabelecido que a análise cromossômica em microarranjos (CMA), uma técnica da citogenômica para a detecção de SNP em um amplo espectro de regiões do genoma humano. O uso de ensaios citogenômicos baseados em microarranjo de alta densidade de DNA permite identificar as variações em SNPs e, consequentemente, genotipá-los. No presente estudo, usando o ensaio do GeneChip® HD™ 750 K foi possível estabelecer os genótipos dos SNPs em uma população nascida de progenitores expostos à radiação ionizante de césio-137. Os desvios mendelianos da linha germinativa foram usados para se estimar a frequência de mutações induzidas pela exposição parental na sua prole. O grupo exposto foi constituído por 11 famílias, dos quais pelo menos um dos progenitores foi diretamente exposto à radiação ionizante de Césio-137, incluindo um total de 37 indivíduos (11 casais e 15 filhos nascidos após o acidente). A dose absorvida para os indivíduos expostos variou entre 0,2 a 0,5 Gray. Um grupo de indivíduos não-expostos à radiação ionizante foi usado como controle. Esse grupo foi composto por 15 famílias goianas sem histórico de exposição à RI. Os testes estatísticos utilizados foram: teste de *Shapiro-Wilk*, teste *F*, análise de regressão, clusterização e análise de componente princial. Todas as análises foram realizadas utilizando o pacote estatístico R, com nível de significância de 5% ($p<0,05$). As frequências de FM_{DM} foram estimadas por caso e controle, representando $1,3x10^{-3}$, $0,9x10^{-3}$, respectivamente. Assim, as frequências de desvios mendelianos mostraram diferenças estatisticamente significativas entre os grupos expostos e controle ($p<2x10^{-3}$, $\alpha = 0,5$, *Student-t*). O teste *F* para comparar as variações de duas amostras (Caso e Controle) de populações com distribuição normal mostrou que as variações ($F=4,47$; $\alpha = 0,5$, $p<8x10^{-3}$) entre casos e controles foram significativamente diferentes. Além disso, a progênie de uma população acidentalmente expostos a baixas doses de RI mostrou ~ 1.44x mais de desvios mendelianos (DM) *de novo* que controles saudáveis. Em conclusão, o estudo foi possível gerar a frequência da mutação de linhagem germinativa/geração em DM poderá ser útil para estudar retrospectivamente populações

humanas expostas à RI, utilizando a técnica de achados de desvios mendelianos, foram possíveis identificar a origem dos progenitores, como também o tipo de substituição e informar qual a variante que sofreu a mutação. Portanto, os DM são marcadores potencialmente úteis para discriminar exposição parental à RI.

Palavras-Chaves: *Bioinformática; SNPs; Genotipagem; Desvio Mendeliano; CytoScan 750K™.*

ABSTRACT

In 1987, in the city of Goiânia, a series of unexpected events resulted in a serious radiological accident, generated by Césio-137. The mutagenic effects of ionizing radiation (IR) can lead to accumulation of mutations in children of irradiated parents. It was established that chromosomal microarray analysis (CMA), a cytogenomic technique for the detection of SNP in a wide spectrum of regions of the human genome. The use of cytogenetic assays based on high-density microarray of DNA allows identifying variations in SNPs and, consequently, genotyping them. In this study, using the GeneChip® HD® assay it was possible to establish the genotypes of SNPs in a population born from progenitors exposed to cesium-137 ionizing radiation. Mendelian germline deviations were used to estimate the rate of mutations induced by parental exposure in their offspring. The exposed group consisted of 11 families, of which at least one parent was directly exposed to Cesium-137 ionizing radiation, including a total of 37 individuals (11 couples and 15 children born after the accident). The absorbed dose for exposed individuals ranged from 0.2 to 0.5 Gray. A group of individuals not exposed to ionizing radiation was used as a control. This group consisted of 15 families from Goiás with no history of exposure to IR. The statistical tests used were: Shapiro-Wilk test, F test, regression analysis, clustering, and the main component analysis. All analyses were performed using the statistical package R, with a significance level of 5% ($p<0.05$). FM_{DM} frequencies were estimated by case and control, representing $1.3x10^{-3}$, $0.9x10^{-3}$, respectively. Thus, the frequencies of FM_{DM} showed statistically significant differences between the exposed and control groups ($p<2x10^{-3}$, $\alpha = 0.5$, Student-t). The F test to compare the variations of two samples (Case and Control) from populations with normal distribution showed that the variations ($F=4.47$; $\alpha = 0.5$, $p<8x10^{-3}$) between cases and controls were significantly different. In addition, the progeny of a population accidentally exposed to low doses of IR showed ~ 1.44x more *de novo* Mendelian deviations (MD) than healthy controls. In conclusion, the frequency of germ line/generation MD mutation may be useful to study human populations exposed to IR retrospectively, using the Mendelian deviation findings technique, it was possible to identify the origin of the parents, as well as the type of substitution and inform which variant suffered the mutation. Therefore, DM are potentially useful markers to discriminate parental exposure to IR.

Keywords: *Bioinformatics; SNPs; Genotyping; Mendelian Deviation; CytoScan 750K™.*

1. INTRODUÇÃO

Com os avanços nas tecnologias de análise genômica grandes volumes de dados de sequências de nucleotídeos são produzidos, incluindo a possibilidade de se identificar e catalogar milhares de polimorfismos de base única (SNP – *do inglês, single nucleotide polymorphism*) com alto grau de precisão. As variações dos SNP são importantes para se determinar as relações genotípicas e fenotípicas inter e intraespecíficas e inter e intrapopulacionais, bem como a identificação de variantes relacionadas às doenças humanas e animais (SHAH e KUSIAK, 2004).

As mutações são os eventos subjacentes aos polimorfismos e, portanto, geram ampla variação genética e são a principal força motriz da evolução. Assim, examinar as taxas e os tipos de mutação é essencial para se compreender as bases genéticas da anatomofisiologia e a evolução dos organismos (TATSUMOTO *et al.,* 2017).

Existem vários métodos para a detecção das variantes estruturais dos genomas (CARTER, 2007; KORBEL *et al.,* 2007). Vários algoritmos de classificação podem ser aplicados aos dados obtidos pela técnica de microarranjo para o desenvolvimento de métodos que possam predizer a ocorrência de uma doença (MUDUNURI *et al.,* 2009). Entretanto, Os chips de genotipagens em SNPs que mais se destacam no mercado são oferecidos pela empresa Thermo Fisher e Illumina (KENNEDY *et al.,* 2003; PEIFFER *et al.,* 2006).

Os algoritmos baseados em modelos estatísticos e métodos não-paramétricos, são aplicados para detectar os SNPs, utilizando as intensidades de fluorescência dos marcadores. No entanto, esses algoritmos necessitam de especificidade para detectar desvios mendelianos, devido à elevada taxa de falsos positivos com base nos valores de intensidade de fluorescência. Para isso, testes de associações são aplicadas para identificar os desvios mendelianos (DM) em SNPs, como também mostrar suas origens parentais (XU *et al.,* 2011).

As matrizes da sonda GeneChip® Cytoscan HD (Thermo Fisher Scientific, Massachusetts, EUA) são feitas usando síntese combinatória direcionada à luz com padrão especial e contém até centenas de milhares de oligonucleotídeos diferentes aderidas a uma pequena superfície de vidro, montados em um *chip*, denominado *array*. Sendo que estes *arrays* genômicos permitem examinar centenas de milhares de sequências-alvo em uma única hibridação e detectam perdas e ganhos de segmentos de DNA cerca de duas ordens

de magnitude menores do que o que pode ser observado ao microscópio (Manual do *Enterprise Affymetrix® Chromosome Analysis Suite 2.0 TM Software User*).

Este conjunto de informações estão contempladas SNPs de alta densidade, com densidade de sonda no CytoScan HD© suficiente para maior que 99% em sensibilidade e maior que 99% para especificidade, sendo possível a visualização de padrão de desequilíbrio alélico, identificação de mutação genômica, verificação de consistência em trio e análise de paternidade, portanto, contribuindo com as interpretações dos resultados genômicos (ZAHIR e MARRA, 2015).

Estudos recentes corroboram para a existência de evidência para a hipótese de mutação *de novo* e a sua relação com a hipótese tardia da paternidade (DE KLUIVER *et al.,* 2017). Foi demonstrado que a idade paterna avançada afeta independentemente todo o espectro da fertilidade masculina, conforme avaliado pela redução na qualidade e na fertilização dos espermatozoides, sendo ela assistida ou não assistida. Sendo que, estudo encontraram evidências crescentes também que sugerem aumento do risco de doenças pediátricas e em adultos, variando de câncer a características comportamentais (CONTI e EISENBERG, 2016).

2. REVISÃO BIBLIOGRÁFICA

2.1. Bioinformática

A partir do final da década de 1980, o termo "bioinformática" tem sido utilizado frequentemente para se referir a métodos computacionais para análise comparativa de dados do genoma. No entanto, o termo foi originalmente mais amplamente definido como o estudo de processos para modelar sistemas biológicos (HOGEWEG, 2011; HOGEWEG & HESPER, 1978; HOGEWEG, 1978).

Com o avanço da pesquisa em inteligência artificial aplicada a novas representações de sistemas de processamento de informações, geralmente inspiradas em sistemas biológicos, por exemplo, modelos de redes neurais para aprendizado e reconhecimento de padrões (MINSKY, 1969; ROSENBLATT, 1961), algoritmos genéticos para otimização em processamento paralelo semi-independente (PAPERT, 1990; ABELSON e DISESSA, 1986; HEWITT, 1977; HOLLAND, 1975), demonstra o poder de uma abordagem autocentrada individualmente para gerar e/ou entender mais estruturas globais (HOGEWEG, 2011).

Desta forma, a reintrodução de ideias computacionais com inspiração biológica contribuiu com o entendimento dos sistemas biológicos como sistemas de processamento de informações. Em particular, um foco na interação local levando a fenômenos emergentes em várias escalas parecia estar ausente na maioria dos modelos biológicos. Contudo, a combinação das análises de padrões, modelagem dinâmica e aplicação busca o desafio de desvendar a geração de padrões e processos de informática em sistemas bióticos em várias escalas (HOGEWEG, 2011).

Com isso, impulsionado pelo aumento exponencial dos dados de sequenciamento, o termo bioinformática passou a significar o desenvolvimento e o uso de métodos computacionais para gerenciamento e análise de dados de dados aplicados para estudos na estrutura de proteínas, previsão de função baseada em homologia e filogenia. No entanto, as ideias valiosas obtidas com os grandes projetos de sequenciamento e a análise bioinformática relacionada para desvendar a função e a evolução tornou-se o “tronco da bioinformática" (HOGEWEG, 2011).

Com o sequenciamento do genoma humano obtivemos uma riqueza de informações detalhando milhões de variações genéticas entre indivíduos. Portanto, abre-se novas

oportunidades para identificar a predisposição genética e entender as causas de doenças comuns. Estima-se que 90% das variações genéticas em humanos sejam devidas a prevalência de SNPs (KOLKMAN, *et al.,* 2007; COLLINS *et al.,* 1997).

Em geral a bioinformática está preocupada em fazer as perguntas certas, gerando e testando hipóteses, além de organizar e interpretar uma imensa quantidade de dados para detectar padrões biológicos (BARNES e GRAY, 2003).

2.2. Matrizes de Genotipagem de Polimorfismo em Base Única

A variação estrutural no genoma humano tem sido intensamente estudada (DE SMITH *et al.,* 2007; REDON *et al.,* 2006; TUZUN *et al.,* 2005; IAFRATE *et al.,* 2004; SEBAT *et al.,* 2004). Existem vários métodos para a detecção dessas variantes estruturais (CARTER, 2007; KORBEL *et al.,* 2007), desta forma, foram aplicados métodos para interpretar resultados de SNP, com base nas análises de microarranjos de DNA em ensaios citogenéticos, utilizando o *kit* de reagentes da *CytoScan* e a matriz *CytoScan* HD™ (*high-density*) para os achados de genotipagem, seguindo-se os protocolos do fabricante.

Os microarranjos são uma ferramenta poderosa de diagnóstico que pode gerar uma gama de informações consideradas relevantes para mapear genes associados a diversas doenças humanas, incluindo câncer (GORLOV *et al.,* 2014). Vários algoritmos de classificação podem ser aplicados na técnica de microarranjo para o desenvolvimento de métodos que possam predizer a variação genômica putativamente associada a uma doença. No entanto, a precisão de tais métodos difere de acordo com o algoritmo de classificação aplicada, e identificar o melhor algoritmo de classificação se torna um grande desafio (MUDUNURI *et al.,* 2009).

A pesquisa genômica pode levar a descobertas de variação de sequência em genes humanos, que está amplamente confinada a SNPs e é valiosa em testes de associação com doenças comuns e características farmacogenéticas (HALUSHKA *et al.,* 1999).

Os chips de genotipagens em SNPs que mais se destacam no mercado são oferecidos pela Affymetrix e Illumina. Ambas as empresas comercializam microarranjos concorrentes e continuam a oferecer maior cobertura para detectar eventos relacionados na análise de número de cópias e ensaios de SNPs, simultaneamente. Sendo que a técnica de ensaio para os microarranjos difere, contudo, a saída de intensidade de sinal das duas

plataformas e apresenta problemas de análise e interpretação semelhantes (PEIFFER *et al.,* 2006; KENNEDY *et al.,* 2003).

Portanto, a base de dados de SNPs são tipicamente padronizados em relação a uma população de referência para reduzir o efeito de fatores, incluindo a variação entre as matrizes e efeitos de hibridação específicos da sonda. Ao fazer isso, as rotinas de normalização assumem implicitamente que todos os membros (ou a grande maioria) da população de referência têm o mesmo número de cópias, no entanto, em locais de SNP em comum, gerando uma suposição de posição na sonda (WINCHESTER *et al.,* 2009).

Contudo, esta base de dados de SNPs, além dos genes, somente foi possível devido à posição dos marcadores por estarem posicionados de acordo com a mesma construção do genoma (KOED *et al.,* 2005).

XU *et al* (2011) afirma que as matrizes de genotipagem foram desenvolvidas para caracterizar os SNPs. Os algoritmos baseados em modelos estatísticos e métodos conhecidos como não-paramétricos, são aplicados para detectar SNP, utilizando as intensidades de marcadores. No entanto, esses algoritmos necessitam de especificidade para detectar DM, devido à elevada taxa de falsos positivos com base nos valores de intensidade. Para isso, o uso de testes de associações é desenvolvido para identificar os DM em SNPs, como também mostrar suas origens parentais.

De acordo com CAETANO (2009), a identificação de SNP's distribuídos aleatoriamente pelo genoma está distribuída por meio do alinhamento de uma sequência de fragmento aleatório do genoma com uma sequência referência.

Entretanto, existe outro método de identificação, sendo por meio de sequenciamento direto de fragmentos específicos do genoma amplificados por PCR, e subsequente alinhamento e comparação das sequências, desta forma aplica a mineração de SNP's em uma região de interesse (CAETANO, 2009).

2.3. Tecnologia Affymetrix – Genotipagem por SNP

As matrizes da sonda GeneChip® contêm até centenas de milhares de oligonucleotídeos diferentes em uma pequena superfície de vidro. As matrizes foram projetadas e usadas para medições quantitativas e altamente paralelas da expressão gênica, para descobrir *loci* polimórficos e detectar a presença de milhares de alelos alternativos (Manual do *Enterprise Affymetrix Chromosome Analysis Suite 2.0 Software User*).

Os atuais métodos de fabricação comercial em larga escala permitem que aproximadamente mais de 750.000 sondas SNP fornecem amplos recursos para definir variações de cópia neutra com confiança (ZAHIR e MARRA, 2015) e aproximadamente 1,9 milhões de marcadores de CNV (do inglês, *copy number variations*), não polimórficos, totalizando aproximadamente 2,6 milhões de marcadores.

Estes microarranjos genômicos permitem examinar centenas de milhares de sequências-alvo em uma única hibridação e detectam perdas e ganhos de segmentos de DNA, cerca de duas ordens de magnitude menores do que o que pode ser observado ao microscópio.

O protocolo CytoScan™ inclui as ferramentas para auxílio na identificação dos marcadores SNPs, sendo que o chip utilizado é o CytoScan HD™ Array-Affy, sendo possível, em um único chip, localizar e identificar os marcadores SNPs e CNVs.

Técnicas de genotipagem por microarranjos tem sido amplamente utilizado como métodos convencionais para detectar LOH (do inglês, *Loss of heterozygosity*) e CNVs (PINKEL *et al.,* 2010). Além de variação estrutural genotípica, sendo uma importante classe de variabilidade genética mendeliana como causa de doenças hereditárias comuns e câncer (WAIN *et al.,* 2009; STANKIEWICZ e LUPSKI, 2010).

Essas técnicas são consideradas eficientes para detectar tais alterações no genoma devido à distribuição relativamente uniforme das sondas (VASSON *et al.,* 2013). Muitos estudos de alto impacto foram baseados em resultados derivados do método microarranjos com alto desempenho (LYBÆK, *et al.,* 2009; VERMEESCH *et al.,* 2007).

Na visão geral de alto nível de como as chamadas de número de cópias e genotipagem de SNP são geradas pelo software ChAS®. O fluxograma começa com as adaptação e correção do fragmento, inclui normalização e dimensionamento, análise das sondas de SNP, em sequência monta o resumo do sinal, para realizar o cálculo de sinal alélico, constituindo a genotipagem, posteriormente identifica a correção da diferença alélica GC (do inglês, *content changes*) para assim, rastrear os picos nos alelos computacionalmente, definindo a detecção de LOH. Conforme representado na Figura 1 abaixo.

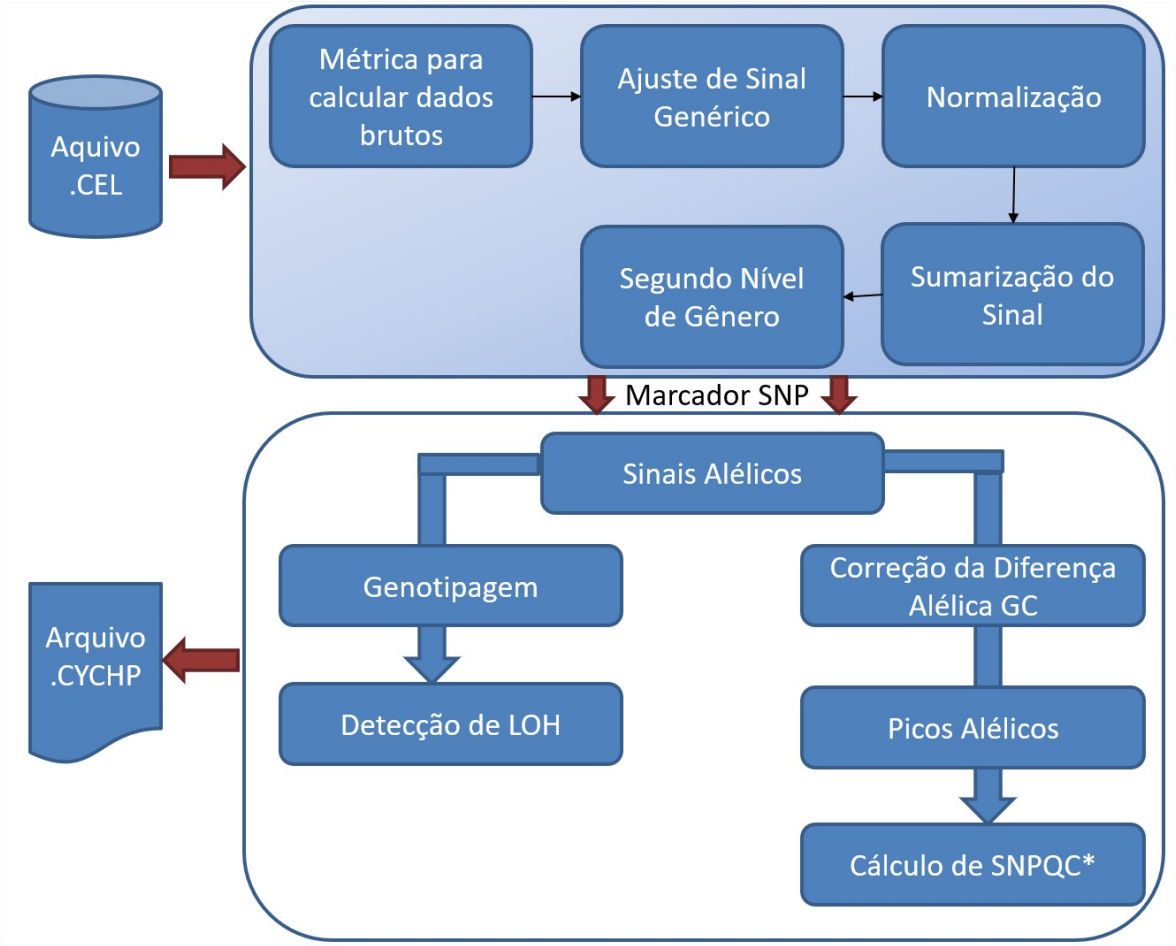


Figura 1. Representando uma visão geral, em forma de fluxograma, da análise do pipeline aplicada ao protocolo CytoScan HD ™. *SNPQC é uma métrica de qualidade para medir os alelos dos genótipos dentro dos microarranjos. (Manual do *Enterprise Affymetrix® Chromosome Analysis Suite 2.0 TM Software User*)

A identificação de recursos e extração de sinal nos microarranjos são digitalizados pelo software GeneChip® Command Console® (AGCC®, Affymetrix, USA). O AGCC® alinha uma grade no arquivo DAT (a imagem digitalizada original) para identificar cada espaço de microarranjo e calcular o sinal de cada recurso. Esse processo usa o arquivo DAT, contendo o sinal bruto, e cria um arquivo CEL, que contém uma única intensidade de sinal para cada recurso. O arquivo .CEL é usado para todas as análises posteriores.

A genotipagem por microarranjo CytoScan HD™ é realizado usando o algoritmo BRLMM-P-Plus. Este algoritmo possui um bom desempenho, além de requerer a presença de sondas com incompatibilidade na matriz para criar genótipos de origem. Sendo que, estes genótipos origem são gerados a partir de propriedades a base de clusterização dos dados (Affymetrix, 2007).

A abordagem do algoritmo BRLMM-P-Plus compreende os dados brutos (.CEL), que com o normalização geram sondas normalizadas, inclui o resumo dos dados dos alelos para a montagem dos sinais (valores estimados), cujo a realização do espaço clusterizado é possível realizar a transformação dos dados com apresentaram contraste, sinalizando as chamadas genotípicas, que apresentam os resultados de genótipos e seus respectivos valores de confiança para cada marcador SNP. Abaixo, representamos o fluxograma do algoritmo BRLMM-P-Plus.

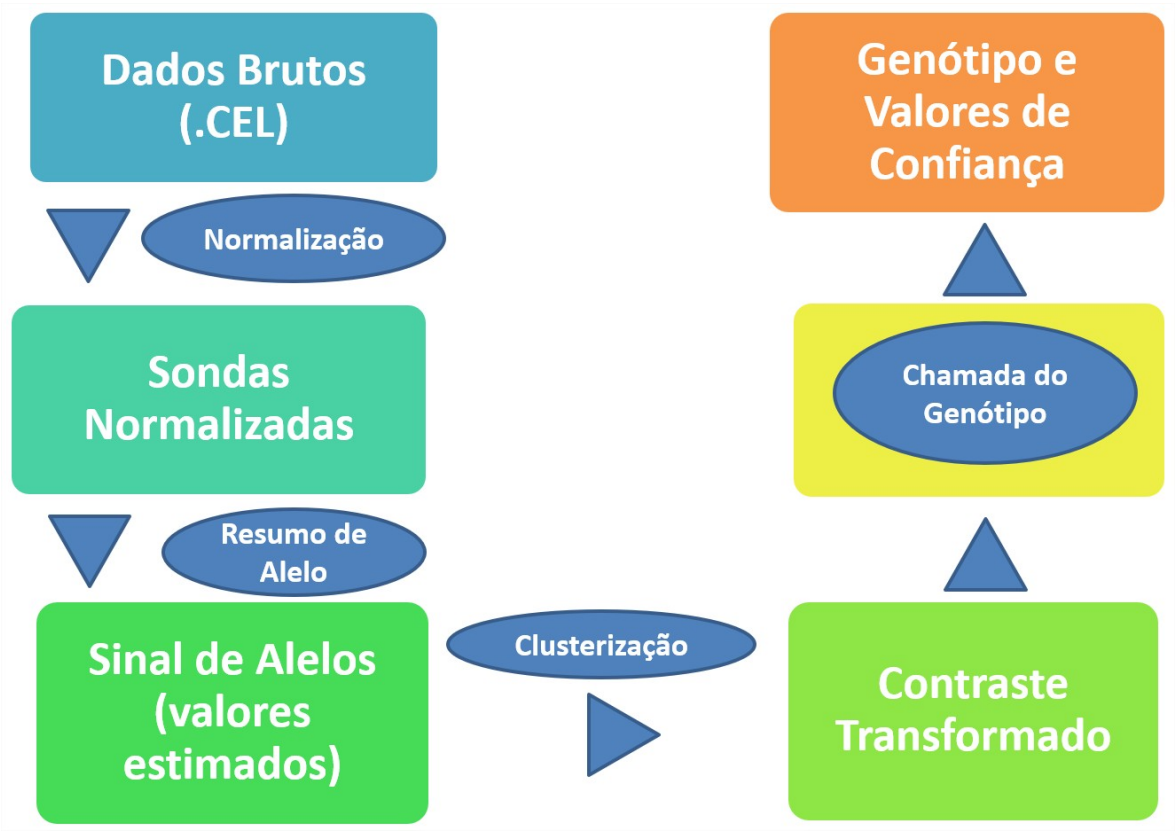


Figura 2. Visão geral da representação das etapas construídas pelo algoritmo BRLMM-P-Plus (Affymetrix, 2007).

2.4. Similaridade entre os softwares de mercado: Affymetrix e Illumina

Durante os últimos anos, as matrizes de genotipagem de SNPs de alta densidade facilitaram estudos de associação em todo genoma (GWAS, do inglês, *Genome-Wide Association Study*) que identificaram com sucesso variantes genéticas comuns associadas a uma variedade de fenótipos. No entanto, cada uma das variantes genéticas identificadas explica apenas uma parte da contribuição genética subjacente à característica fenotípica

estudada. Além disso, a discordância observada nos resultados entre o GWAS independente indica o potencial de erros do tipo I e II (HONG *et al.*, 2012).

É necessária alta confiabilidade da tecnologia de genotipagem para ter confiança no uso de dados SNP e na interpretação dos resultados do GWAS. Estudos demonstram uma confiabilidade técnica das plataformas de genotipagem atualmente disponíveis, porém indicam a importância de incorporar algumas técnicas para a genotipagem do QC[1] (do inglês, *Quality Control*), a fim de melhorar a confiabilidade dos resultados de genotipagem em SNP. O impacto de genótipos discordantes pode explicar, pelo menos em parte, a irreprodutibilidade de alguns achados de genotipagem em SNP de alta densidade quando o tamanho do efeito e as frequências de alelos menores são baixas (MAREES *et al*., 2018).

Estudos demonstram que quando aumenta o tamanho da amostra, para maior obtenção do poder estatístico, consequentemente o efeito de uma taxa de erro de genotipagem equivalente diminui (KERKHOF *et al*., 2010; PETERSEN *et al*., 2010; AZZATO *et al*., 2010; KNAUFF *et al*., 2009; LYSSENKO e GROOP, 2009; LANDI *et al*., 2009; PFEUFER *et al*., 2007). Sendo que, é possível que duas pessoas diferentes produzam dois resultados ligeiramente diferentes utilizando o mesmo protocolo do fabricante. Estes protocolos são aplicados por meio de microarranjos com alta densidade são focados em variantes raras, sendo que o critério de seleção para que os SNPs sejam detectados, em pelo menos, vários conjuntos de dados de sequenciamento (GUO, *et al*., 2014).

Estudos revelam que os algoritmos aplicados nos protocolos com matrizes de genotipagem de SNP com alta densidade, geralmente são altamente concordantes para a maioria dos SNPs, principalmente para os dois fornecedores mais comuns de SNP em microarranjos: Illumina e Affymetrix (ECKEL-PASSOW *et al.,* 2011).

Um estudo publicado sobre a comparação das técnicas de reprodutibilidade das fabricantes Affymetrix e Illumina no uso das matrizes de genotipagem de SNP concluiu principalmente que as tecnologias apresentaram resultados satisfatórios e razoável tratando de concordâncias nos genótipos, sendo que, um número reduzido de discordância genotípica pode gerar associações falsas, especialmente para marcadores genéticos com

1 Uma etapa essencial para utilizar o GWAS é o uso de QC apropriado. Sem um controle de qualidade extensivo, o GWAS não gera resultados confiáveis, pois os dados brutos do genótipo são inerentemente imperfeitos. Os erros nos dados podem surgir por várias razões, por exemplo, devido à baixa qualidade das amostras de DNA, à hibridação do DNA com o chip, às sondas de genótipo pode possuir desempenho insatisfatório e às misturas ou contaminação da amostra.

baixa frequência alélica. Para estes casos, o estudo sugere um aprofundamento mais cuidadoso na replicação para confirmar estas associações (HONG *et al.*, 2012).

2.5. Software ChAS Affymetrix

Utiliza-se o software *Affymetrix Chromosome Analysis Suite*® (ChAS) para analisar dados do microarranjo CytoScanHD™ e discutir estratégias para um projeto experimental e para determinar a patogenicidade das CNVs identificadas. Com a aplicação do ChAS é possível mapear as intensidades de cada alelo que são usadas para inferir estados de número de cópias, genótipos e intensidades de alelos. Estas informações formam a base de todas as demais análises (ZAHIR e MARRA, 2015). É possível usar o software ChAS para detectar eventos de aberrações altamente complexos que são frequentemente encontrados em tumores de câncer. Dados gerados pelo ChAS podem ser exportados para outras ferramentas de visualização com melhor rendimento diagnóstico (AMBROS *et al.,* 2014). A alta densidade da sonda SNP da plataforma CytoScan é útil para decifrar eventos complexos, pois os dados polimórficos da sonda que podem ser usados para reconstruir o cromossomo de origem.

Assim, dada uma sequência de referência, pode ser projetada uma matriz de sondas de DNA que consiste em uma coleção altamente densa de sondas complementares, praticamente sem restrições nos parâmetros de projeto. Sendo que, a quantidade de informação de ácido nucléico codificada na matriz na forma de sondas diferentes é limitada apenas pelo tamanho físico da matriz e pela resolução eficiente litográfica (LIPSHUTZ *et al.,* 1999).

Para RODIG *et al.,* (2010), os dados extraídos do array de SNP da fabricante Affymetrix podem ser analisados com software proprietário especialmente projetado. No console de genotipagem, as amostras são agrupadas em dentro dos limites (boa amostra) e fora dos limites (amostras com ruídos) após verificações iniciais de controle de qualidade, permitindo ao usuário investigar e analisar a incompatibilidade ou compatibilidade da sonda e do agrupamento dos SNP individualmente.

A recomendação recente é que a análise clínica de análise cromossômicos por microarranjo (CMA – do inglês, *Chromosomal Microarranjo Analysis*) tenha resolução suficiente para detectar deleções submicroscópicas e duplicações de 100 kb ou mais (SOUTH *et al.,* 2013; TEODORO *et al.,* 2017).

As plataformas de CMA contém sondas não polimórficas e específicas para CNV,

capazes de interrogar o estado do número de cópias e sondas oligonucleotídicas curtas com 25 pb, que incluem um polimorfismo de nucleotídeo único (SNP) e são polimórficos, ou seja, podem ser usados para inferir o genótipo (ZAHIR e MARRA, 2015).

O CMA realiza a hibridização da amostra do paciente com os marcadores que foram sintetizados no chip. Posteriormente, por meio de uma análise criteriosa do técnico especialista, o resultado desta análise é comparado com os bancos de dados de população saudável e bancos de dados de doenças.

Para TUCKER *et al.,* (2011), as primeiras versões das matrizes de mapeamento Affymetrix GeneChip®™ foram as matrizes de 100 K, 500 K e 750 K. Essas matrizes foram projetadas para avaliar genótipos, mas foram adaptadas para análise de número de cópias.

Na figura abaixo representamos as etapas na utilização do CMA. O especialista deve tomar as decisões quanto à composição das amostras a serem coletadas e quais as amostras serão constituídas como referência. A verificação deve ser realizada por um desses métodos de maior acurácia. A patogenicidade deve ser acompanhada por ensaios de validação e deve-se realizar a correlação de genótipo-fenótipo (ZAHIR e MARRA, 2015).

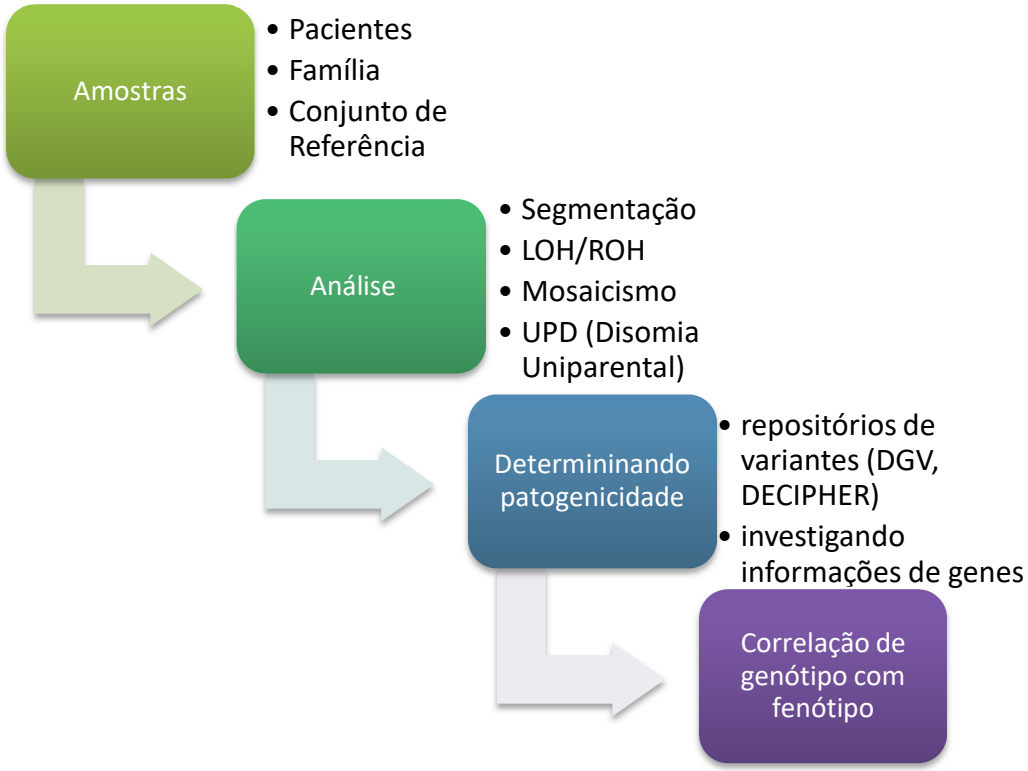


Figura 3. Etapas usando a metodologia CMA que está incorporada na plataforma ChAS® (ZAHIR e MARRA, 2015).

Todo o conjunto de informações contempladas em microarranjos de genotipagem em SNPs de alta densidade, com precisão de genótipo maior que 95%, sendo possível, por meio do software ChAS, a visualização de padrão de desequilíbrio alélico, identificação de contaminação genômica, verificação de consistência em trio e análise de paternidade, portanto, contribuindo com as interpretações dos resultados genômicos.

2.6 Soluções Recursivas

Problemas computacionais contêm soluções de instâncias menores do mesmo problema, assim pode se afirmar que tais problemas têm estrutura recursiva (Feofiloff, 2009). Quando um algoritmo contém uma chamada recursiva, ou seja, chamada interna, seu tempo de execução pode ser descrito por uma recorrência (CORMEN et a. 2009; ERICKSON, 1999; AHO e JOHNSON, 1974).

Como todas as estruturas recursivas, uma recorrência consiste em um ou mais casos para solução do problema (ERICKSON, 1999). As soluções recursivas envolvem as divisões dos cálculos de uma função em mais subfunções igualmente complexas, cuja avaliação pode proporcionalmente ser realizada simultaneamente em processos separados (KOOGE e STONE, 1973). A figura 4 abaixo descreve a forma que as soluções recursivas são tratadas na divisão do processamento.

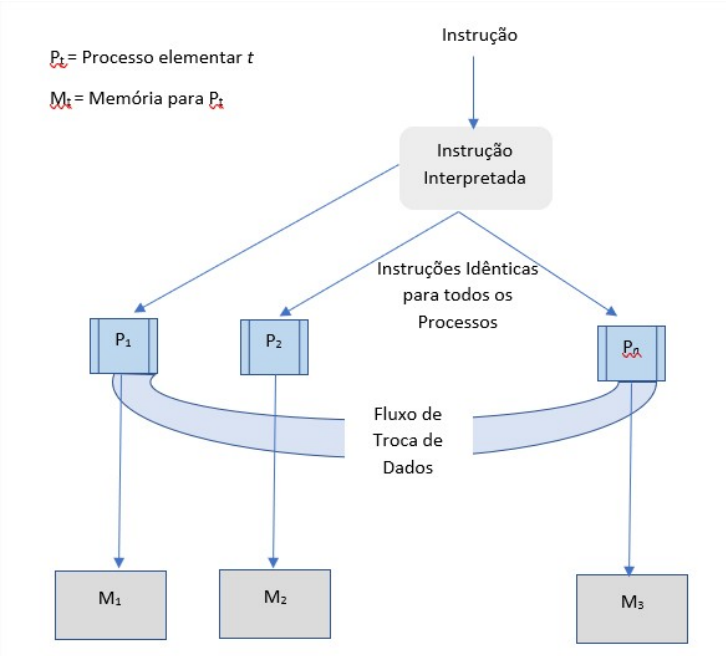


Figura 4. Funcionamento de uma divisão de processamento em um modelo computacional (KOOGE, 1974).

Esboçado a recursividade, em que cada nó do problema representa o custo de *n* subproblemas, sendo que o conjunto de invocações determina uma função recursiva. Este conjunto de invocações denominamos de árvore de recursão (CORMEN *et al.,* 2009). A solução para esta teoria foi o algoritmo conhecido como *mergesort*, sendo que o tempo de execução desse algoritmo é descrito pela **Equação 1** abaixo (ERICKSON, 1999).

$$T(n) = aT\left(\frac{n}{b}\right) + f(n)$$

Onde:

$a \geq 1$, $b \geq 1$ são constantes;
T: tempo de execução;
n: número de problemas;
f(n): função assintoticamente positiva.

O custo de dividir o problema e combinar os resultados dos subproblemas é descrito pela função *f*(n). O valor de *a* representa os subproblemas e *b* informa o tamanho do subproblema. *T(n)* é o somatório de todos os valores armazenados na recursividade da árvore. Para cada valor de *i,* a enésimo nível da árvore contém a^i nós, para cada valor de $f(a/b^i)$, então, segue abaixo a equação 2.

Equação 2

$$T(n) = \sum_{i=0}^{L} a^i f(n/b^i)$$

, **onde** L representa a profundidade da árvore, sendo que $L = \log_b n$ (ERICKSON, 1999).

A figura 5 abaixo representa a recursividade *mergesort,* descrevendo o somatório de níveis de subproblemas em uma série geométrica, aplicada para pequenos ou grandes escalas.

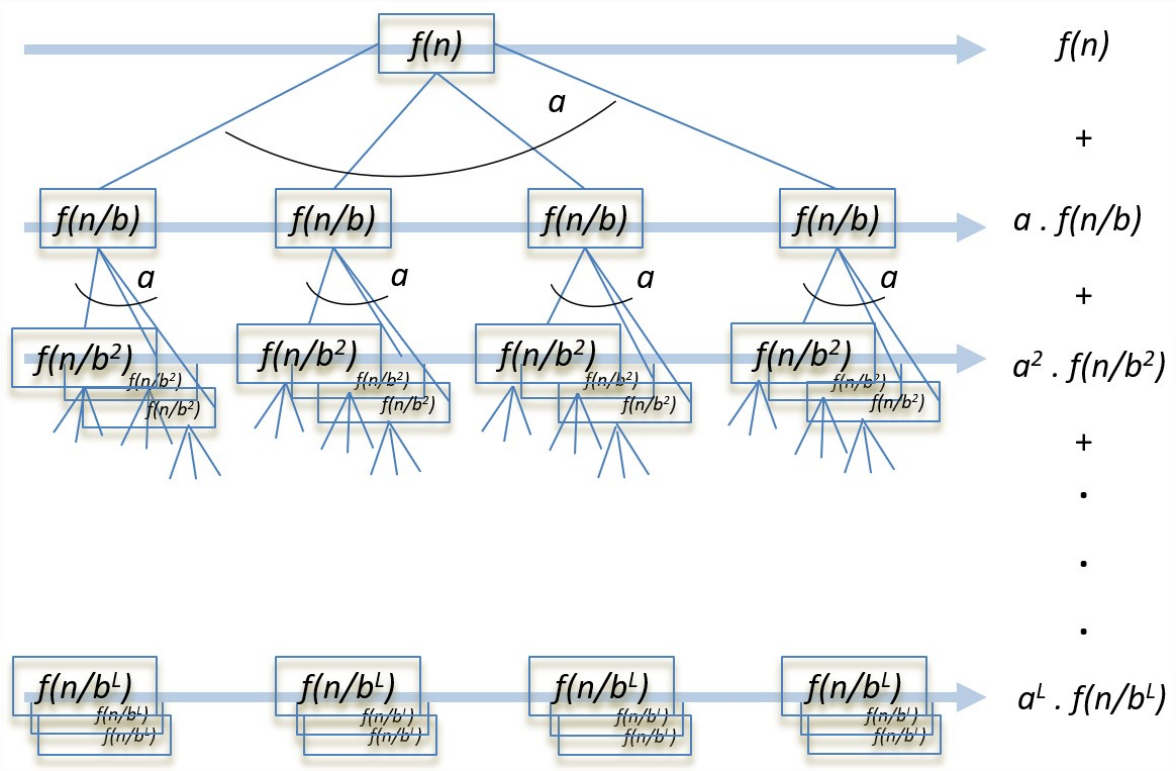


Figura 5. Descrevendo uma série geométrica para *n* recorrências (ERICKSON, 1999).

2.7 Cálculo Proposicional

Cálculos proposicionais consistem na interpretação de fórmulas, ou seja, atribuição dos valores-verdade (verdadeiro ou falso) às formulas atômicas, por exemplo: (pvq) → (p^q), esta fórmula possui 2 componentes atômicos, assim, teremos 2^2(nº de componentes atômicos), ou seja, a fórmula geral pode ser representada por 2^n, onde n é o número de componentes atômicos (MONARD *et al.,* 1992).

Assim, com base nas proposições (por exemplo: a prole herda a mutação de seus progenitores), cria-se argumentos que são formadas por premissas iniciais, e que resultam em uma conclusão.

Um argumento é considerado válido quando verificamos que a conclusão é uma consequência obrigatória das premissas, neste caso, a nossa argumentação é válida quando for possível concluir que a premissa "prole herda a mutação de seus progenitores", pode assumir como uma sentença verdadeira ou falsa (CARVALHO e CAMPOS, 2010).

Com isso, sejam P1, P2,... Pn (n ≥ 1) e C proposições ou premissas quaisquer, simples ou compostas, denominamos de argumento a sequência finita de proposições P1, P2,... Pn (n ≥ 1) que tem como consequência a proposição C (conclusão). Desta forma,

podemos concluir que argumento é um encadeamento lógico de premissas que implicam em uma conclusão (VILLAR, 2016).

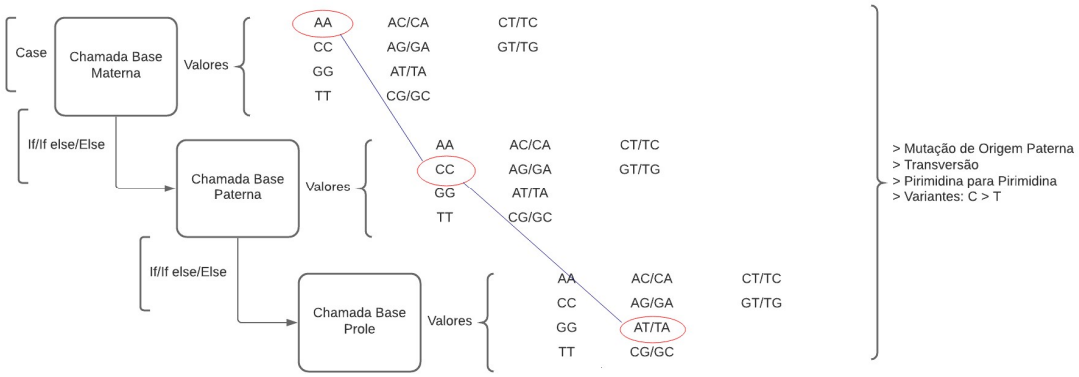


Figura 6. Representa o fluxo lógico para demostrar como as soluções recursivas foram implementadas na solução do problema.

2.8. Histórico sobre o acidente radiológico com Césio-137

Em 1987, na cidade de Goiânia, capital do Estado de Goiás (Brasil), uma série de eventos inesperados resultou em um grave acidente radiológico, levando à exposição radiação ionizante e à contaminação humana, animal, vegetal e ambiental gerado por césio-137 (CRUZ *et al.,* 2010).

Segundo RAMALHO *et al.,* (1988), uma cápsula, utilizada em aparelhos de radioterapia, foi encontrada nas antigas instalações do Instituto Goiano de Radioterapia e continha cerca de 19,26 g de cloreto de césio-137 (137CsCl), que gerou um rastro de contaminação, sem precedentes, em uma dezena de focos espalhados pela região metropolitana de Goiânia (DA CRUZ *et al.,* 1996; DA CRUZ *et al.,* 1994).

O acidente radiológico em Goiânia foi considerado o mais grave do hemisfério ocidental. Durante as 2 semanas seguintes ao acidente, uma área proporcional a 2000 m^2 foi contaminada (DA CRUZ *et al.,* 2008).

Foram identificadas 249 pessoas expostas de forma significativa à radiação ionizante de Césio-137. Para algumas pessoas, a exposição individual resultou da contaminação interna e externa ao sal radioativo, outras foram expostas à energia radiativa emitida pelo decaimento do césio-137. Em alguns casos, observou-se exposição e contaminação individuais. As doses absorvidas de RI durante o acidente variaram de 0 a 7

Gy, que resultou em quatro mortes durante a fase aguda do acidente goiano (IAEA, 1998; DA CRUZ, 1997).

Além da exposição acidental de parte da população à RI emitida pelo radionuclídeo, exposição ocupacional também foi registrada para os membros do corpo de Bombeiros e da Polícia Militar, envolvidos na remoção dos rejeitos radioativos, lavagem de asfalto e isolamento dos locais atingidos, e também de outros profissionais envolvidos com atenção e o cuidado aos radiacidentados. A maioria desses grupos de indivíduos afetados receberam exposições durante o período prolongado e de corpo inteiro, o que dificultou a estimativa de dose total absorvida (RAMALHO *et al.,* 1998).

Segundo OKUNO (2013), todo o rejeito foi armazenado em uma cidade no entorno de Goiânia, Abadia de Goiás, situada a 23 km do centro de Goiânia. Neste local, foram construídas seis plataformas cada uma com 60 ×18 m^2, sobre as quais foram colocados os rejeitos armazenados em 4.223 tambores de 200 L cada, 1.347 caixas metálicas de 1,7 m^3 cada, 10 contêineres marítimos de 32 m^3 cada e seis embalagens especiais construídas com concreto armado com 20 cm de espessura.

Desde o acidente, até os dias de hoje, têm sido realizados vários estudos sobre a saúde genética dos radioacidentados goianos. Um dos primeiros testes de biomonitoramento das populações expostas à radiação ionizante (RI) do Césio-137 foi o teste de micronúcleo, que relatou um aumento na frequência de micronúcleos das pessoas envolvidas direta ou indiretamente no acidente (DA CRUZ *et al.,* 1994). Outros estudos que se destacaram foram o de análise dos níveis de mutação *in vivo* utilizando o parâmetro da expansão clonal de linfócito T contendo mutações no gene HPRT dos indivíduos expostos à radiação (DA CRUZ *et al.,* 1997), análise de aberrações cromossômicas, avaliação de marcadores sorológicos de autoimunidade e análise de mutações germinativas usando marcadores STR de indivíduos acidentalmente e ocupacionalmente expostos à RI (DA CRUZ *et al.,* 1997; DA CRUZ *et al.,* 1996; DA CRUZ *et al.,* 1994). Estudos mais recentes investigaram as análises de CNV's *de novo* como biomarcadores de exposição radiativa (COSTA, *et al.,* 2018).

2.9. Exposição dos sistemas biológicos à radiação ionizante

A dose de uma exposição à radiação ionizante é uma medida com base na projeção que representa o efeito biológico prejudicial de forma geral. O cálculo pondera-se pelas

concentrações de energia depositada em cada órgão a partir de uma exposição à radiação, por meio de uso de parâmetros que refletem o tipo de radiação e o potencial de alterações mutagênicas relacionadas à radiação em cada órgão ou tecido com um valor de referência (FAZEL *et al.,* 2009).

Sabe-se que a exposição de uma ampla variedade de células à irradiação ionizante (RI) (X- ou γ-) resulta em um atraso de divisão das células, atrasando a progressão normal através do ciclo celular (ILIAKIS 2003; BERNHARD *et al.,* 1995; MAITY *et al.,* 1994). Estes atrasos foram inicialmente interpretados como respostas celulares passivas resultantes da indução por RI durante o dano provocado no DNA (ILIAKIS *et al.,* 2003).

Entretanto, esses estudos iniciais forneceram evidências circunstanciais de que os atrasos refletem a indução de processos celulares, em que a célula irradiada deve se adaptar com o dano induzido, facilitando de alguma forma o reparo do DNA e consequentemente o dano celular (KAO *et al.,* 2007; LÜCKE-HUHLE, 1982; TOBEY 1975; WALTERS *et al.,* 1974). A exposição leva a uma resposta aos danos no DNA para permitir o reparo das quebras da fita de DNA antes da divisão celular acontecer (METTLER *et al.,* 2008).

Por meio de intensos e diversificados estudo genéticos foram identificadas complexas redes de genes que cooperam para retardar a progressão normal ao longo do ciclo, assim que os danos são registrados no genoma. Desta forma, é conhecido que os mecanismos de reparo do genoma, por meio do atraso na progressão do ciclo celular é apenas uma manifestação, capturada pelo ponto de verificação de danos ao DNA, de um processo fisiológico inerente ao bom funcionamento e a homeostase adequada das células (ZHOU e ELLEDGE, 2000; ELLEDGE, 1996).

No contexto acima mencionado, a RI é um agente mutagênico amplamente estudado, a exposição celular à energia radioativa resulta em diferentes tipos de lesões no DNA, que vão desde as mudanças em nucleotídeos, até o rompimento de fita da cadeia dupla de DNA. Portanto, prevalece a relevância biológica de cinco tipos de lesões no DNA induzidos por RI para estudos cinéticos, sendo que foram realizados em células eucarióticas por vários laboratórios. Estas lesões foram identificadas em ligações cruzadas de DNA-proteína, danos às bases, quebras de cadeia simples e quebras de cadeia dupla (ADEWOYE *et al.,* 2015; FRANKENBERG-SCHWAGER, 1990).

Os efeitos mutagênicos da RI na linha germinativa são particularmente preocupantes, pois levam ao acúmulo de mutações extras na prole de progenitores irradiados (ADEWOYE *et al.,* 2015). Com isso, esses efeitos mutagênicos influenciam diretamente na linhagem germinativa, assim, se tornam particularmente preocupantes por carregar o acúmulo de mutações adicionais em filhos de pais irradiados. Apesar dos numerosos esforços, pouco se sabe sobre os efeitos genéticos da exposição à radiação em seres humanos e a maior parte da evidência consolidada advém de extrapolações a partir da indução de mutação germinativa *in vivo* em mamíferos, frequentemente ratos e camundongos (NAKAMURA *et al.,* 2013; UNSCEAR, 2001).

2.10 Efeitos das idades parentais nas mutações

Estima-se que a taxa de mutação e identificação de modelos de mutações são importantes para compreender o mecanismo molecular de uma condição fisiológica de um organismo e da história da evolução das espécies (TATSUMOTO *et al.,* 2017).

As mutações germinativas são fontes de todas as adaptações evolutivas e doenças hereditárias, caracterizando as propriedades e as taxas de como os cruzamentos individuais são fundamentais para a genética humana (SÉGUREL *et al.,* 2014).

Estudos recentes corroboram para a existência de evidência para a hipótese de mutação *de novo*, buscando relação com a hipótese tardia da paternidade (DE KLUIVER *et al.,* 2017) Sendo que as mutações *de novo* ocorrem espontaneamente na linha germinativa masculina durante as divisões de células-tronco espermatogonais (JÓNSSON *et al.,* 2017) e se propagam em clones sucessivos de espermatócitos (CROW, 2000; PENROSE, 1955). Tais mutações de novo na linha germinativa masculina ocorrem com mais frequência com o aumento da idade paterna e a hipótese de aumentar a morbidade na prole (DE KLUIVER *et al.,* 2017).

Ainda DE KLUIVER *et al.,* (2017), afirma que existe uma crescente percepção de que, independentemente da idade materna, a idade paterna avançada na gravidez está associada à morbidade dos filhos, incluindo distúrbios psiquiátricos, esta associação é conhecida como efeitos da idade paterna.

Foi demonstrado que a idade paterna avançada afeta independentemente todo o espectro da fertilidade masculina, conforme avaliado pela redução na qualidade e na fertilização dos espermatozoides, sendo ela assistida ou não assistida. Além disso, dados

epidemiológicos sugerem que a idade paterna pode levar a taxas mais altas de resultados adversos ao nascimento e anomalias congênitas (CONTI e EISENBERG, 2016). Na figura 7 abaixo representamos a meiose paterna e materno até o nível de espermatozoide (masculino) e óvulo (feminino).

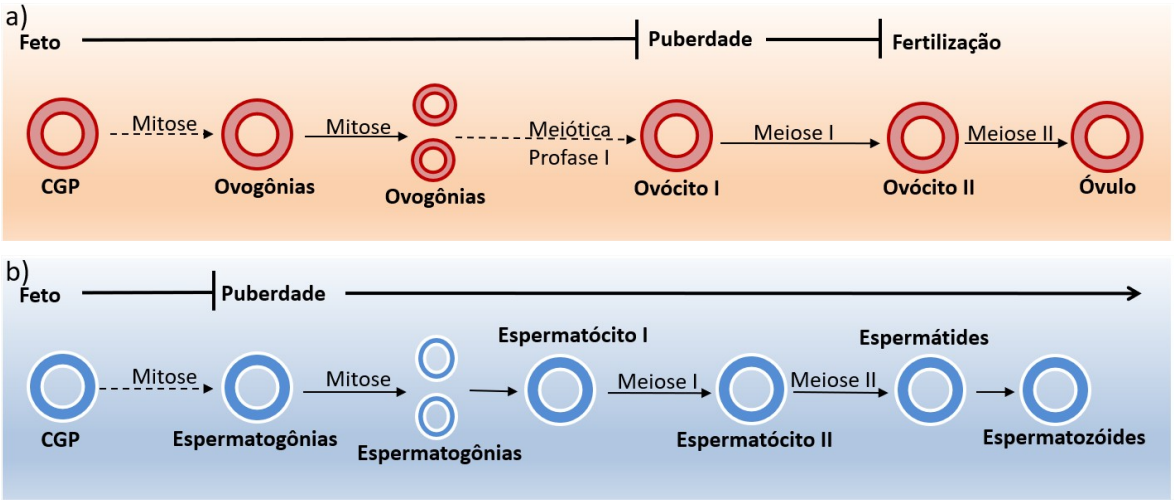


Figura 7. As imagens descrevem o ciclo das células reprodutivas. O ciclo de meiose feminina (a) ocorre na maior parte antes do nascimento, enquanto no ciclo masculina (b) as divisões ocorrem em sua maior parte antes da puberdade (CONTI e EISENBERG, 2016).

Assim, qualquer alteração genética nos espermatozoides, se não for reparada ou eliminada, pode ser transmitida à prole, levando potencialmente a malformações, anomalias cromossômicas e doenças monogênicas. Mutações *de novo* são formadas espontaneamente e tendem a se acumular com maior frequência durante o envelhecimento testicular (CIOPPI *et al.,* 2019).

Apesar do aumento intenso na proporção de divisões de células germinativas masculinas e femininas após o início da espermatogênese, mesmo os pais jovens contribuem com três vezes mais mutações do que as mães jovens, e essa proporção quase não aumenta com a idade dos pais. Com isso, prevalece uma contribuição substancial de mutações induzidas por danos. Esses achados revelam papéis subestimados de danos ao DNA e idade materna na gênese das mutações da linha germinativa humana. Além disso, foram encontradas evidências de que uma fração substancial das mutações não são replicadas na origem e com isso um efeito potencial da idade da mãe no número de mutações que acontecem no início do desenvolvimento do embrião. (GAO *et al*., 2019).

Com isso, estudos revelam que o número de mutações agrupadas aumenta mais rápido com a idade da mãe do que com a do pai, e a extensão genômica dos agrupamentos de mutação materna *de novo* é maior do que a dos paternos (JÓNSSON *et al*., 2017).

2.11 Desequilíbrio de Ligação

As aplicações mais atuais dos dados genômicos envolvem matrizes de polimorfismo de SNP ou dados da sequência do genoma inteiro. Dependendo da diversidade genética das amostras e da densidade das matrizes SNP, pode haver redundância considerável em *lócus* (LARMER *et al.*, 2014) no sentido de que muitos pares de SNPs estão em desequilíbrio de ligação (DL) muito alto ou completo, ou seja, eles têm um valor r^2 (HILL e ROBERTSON, 1968) igual ou próximo de 1. A estatística r^2 favorece a confirmação do histórico de recombinação e indica como os marcadores correlacionam-se com as regiões cromossômicas (DU *et al*., 2007).

O DL tende a se reduzir com o aumento da ocorrência durante a formação dos gametas (AMARAL *et al*., 2008). Para aplicações como previsão genômica, é prática comum remover um SNP de cada par de SNPs com um valor de r^2 igual a 1 (WIGGANS *et al*., 2009). A remoção de *loci* com base em altos níveis de DL em pares é comumente conhecida como remoção de DL. Caso os pares possíveis de *loci* redundantes forem considerados para remoção, o cálculo do DL em pares entre todos os SNPs disponíveis pode não ser computacionalmente viável.

Existem várias ferramentas e bibliotecas que calculam DL em pares entre SNPs (PURCELL *et al*., 2007; BARRETT *et al*., 2005). Essas ferramentas são aplicadas para avaliar o DL em regiões do genoma em que associações significativas foram detectadas em estudos de associação de *GWAS* (PORTO-NETO e KIJAS, 2014; LUO *et al*., 2013; LI *et al*., 2011; DUIJVESTEIJN *et al*., 2010; MEGENS *et al*., 2009; KHATKAR *et al*., 2008; PURCELL *et al*., 2007). Essa avaliação do DL nas regiões do genoma requer o cálculo do DL através de distâncias relativamente curtas no genoma. Por esse motivo, mas também para reduzir os requisitos computacionais gerais, as ferramentas existentes geralmente calculam DL entre pares de SNPs localizados a uma certa distância no genoma, conforme definido pelo usuário. No entanto, para fazer o desbaste de DL, pode ser desejável considerar DL para todas as combinações de pares de loci (CALUS e VANDENPLAS, 2018).

Conforme WIGGANS *et al.* (2009), observaram que SNPs altamente correlacionados deveriam ter MAF (do inglês, *minor allele frequency*) semelhante e, portanto, avaliaram apenas pares de SNPs com diferença no MAF menor que 2,5% das unidades. Esses autores simplesmente consideraram que dois SNPs estão perfeitamente correlacionados se os genótipos forem todos iguais (0–0, 1–1 e 2–2) ou opostos (0–2 e 2–0), enquanto permitindo que 0,5% dos genótipos individuais sejam diferentes dessas regras, para permitir erros de genotipagem.

2.12 PLINK

PLINK, é um pacote de ferramentas, usando a linguagem C/C^{++} de código aberto. Com a utilização do PLINK, grandes conjuntos de dados compreendendo centenas de milhares de marcadores genotipados para milhares de indivíduos podem ser manipulados e analisados em sua totalidade. Além de fornecer ferramentas para tornar as etapas analíticas básicas computacionalmente eficientes, o PLINK também suporta algumas novas abordagens para dados e cobertura do genoma por completo (PURCELL, *et al*., 2007).

Ainda de acordo com Purcell (2007), na análise de associação padrão, uma relação mais distante entre indivíduos que compartilham a mesma doença pode transmitir informações adicionais para o mapeamento genético. Estas análises de associação podem ser capazes de fornecer uma abordagem complementar para estudos de associação de um único SNP.

A ferramenta contempla os principais domínios que atendem o gerenciamento de dados, estatísticas, estratificação populacional, análise de associação, estimativa sobre informações por estado (IBS, do inglês, *identity-by-state*) e informações sobre identidade por descendente (IBD, do inglês, *identity-by-descent*). Em particular, prevalece uma concentração sob as informações geradas a partir de IBS e IBD no contexto de estudos de todo o genoma com base populacional. Essas informações podem ser usadas para detectar e corrigir a estratificação populacional e identificar segmentos cromossômicos estendidos que são compartilhados de forma idêntica por descendência entre indivíduos. A análise dos padrões de compartilhamento tem o potencial de mapear locais de doenças que contêm múltiplas variantes raras em uma análise de ligação populacional (PURCELL *et al*., 2007).

2.13 Medidas de Desiquilíbrio de Ligação para Único Par de SNPs

Por meio do pacote PLINK, aplica-se o comando --ld seguido por dois identificadores SNP, com isso, será gerado os resultados estatísticos do DL para um arquivo LOG, de um único par de SNPs: r^2, D', as frequências estimadas e esperada de haplótipos sob equilíbrio de ligação (EL), ou seja, ocorrendo com mais frequência do que o esperado por acaso. Na tabela 1, temos a representação sobre o DL para os pares de SNPs: rs2840528 e rs7545940.

Informações sobre o DL para os pares de SNPs [rs2840528 rs7545940**]**

R-sq = 0,592 D’ = 0,936

Tabela 1. As fases dos alelos para os haplotipo sob EL dos SNPs rs2840528 rs7545940 representados pela frequência estimada e frequência esperada. As fases dos alelos são GT/AC serão considerados para o desbaste, pois apresentam a maior frequência e estão mais próximas.

HAPLOTIPO	FREQUÊNCIA	FREQ. ESPERADA EL
GC	**0,013**	**0,199**
AC	**0,431**	**0,245**
GT	**0,441**	**0,250**
AT	**0.111**	**0,307**

3 OBJETIVOS

3.1 Objetivo Geral

Aplicar recursos da bioinformática para se estimar a frequência média de desvios mendelianos em SNP, e consequentemente inferir a origem progenitora, tipo de substituição e a variante mutável, de autossomos de um coorte de pessoas concebidas após a exposição parental à radiação ionizante.

3.2 Objetivos Específicos

- Analisar SNP de pai, mãe e filho no intuito de identificar as variáveis das possíveis causas de desvios mendelianos nas proles;
- Estimar a frequência de mutações germinativas na prole de indivíduos expostos acidentalmente à radiação ionizante do Césio-137;
- Propor um modelo, com base em mecanismo de inferências, para identificação da origem progenitora que conduziu o desvio mendeliano aos seus filhos, identificar o tipo de substituição na base do DNA, e indicar a variante que sofreu a mutação;
- Relacionar os efeitos da radiação ionizante como biomarcadores que podem conduzir a existência de desvios mendelianos;

4 METODOLOGIA

4.1 Delineamento do Estudo

O presente estudo foi conduzido no Núcleo de Pesquisas Replicon (NPR) da Escola de Ciências Agrárias e Biológicas (ECAB) da Pontifícia Universidade Católica de Goiás (PUC-Goiás) em parcerias com o Laboratório de Citogenética Humana e Genética Molecular (LAGENE)/ Laboratório de Saúde Pública Dr. Giovanni Cysneiros (LACEN) da Secretaria de Estado da Saúde de Goiás (SES/GO). Análises de bioinformática foram realizadas sob a coorientação do pesquisador Dr. Alexandre Rodrigues Caetano, PhD da Embrapa Recursos Genéticos e Biotecnologia.

Os procedimentos da análise de bioinformática foram aplicados usando-se a versão hg19 do genoma humano disponibilizado pelo *Genome Browser* da *University of California, Santa Clara,* a partir dos dados obtidos pelo GeneChip® CytoScan HD™ da Thermo Fisher para a geração de imagens de dados brutos no nível da sonda, sinais e ruído para a determinação de genótipos e filtros.

Para a identificação das mutações, como também a determinação da sua origem, e a indicação do tipo de substituição, foi aplicado algoritmo que denominamos de SIPO (do inglês, *script inference parental origin)* usando bibliotecas da linguagem de programação *Perl (*do inglês, *Practical Extraction and Report Language*). Assim, SIPO foi escrito na forma de linha de comando, pensado na possiblidade de se trabalhar os dados de SNP, seguindo as observações desta proposta.

Os resultados gerados pelos *scripts* foram armazenados em um banco relacional MySql© - versão 5.0.12, como também instalado um servidor de aplicação para soluções web – Apache® – versão 2.4.29. Neste estudo, o uso de gráficos exploratórios, incluindo, gráficos de caixas, regressões, análise principal de componentes, análise de agrupamento, validação de testes estatísticos foram elaborados no ambiente de programação R© de código aberto em conjunto com as bibliotecas do Biocondutor e CRAN (do inglês, *Comprehensive R Archive Network*), também de código aberto.

Os genótipos vinculados a cada marcador (SNP) dos progenitores foram comparados com os genótipos de seus filhos em uma análise denominada “trio”. Este estudo foi composto por três importantes módulos na construção do deliamento desta

pesquisa. A primeira etapa foi construída no manuseio do software ChAS. O fluxo se iniciou no ponto em que seleciona os arquivos (.CYCHP[2]) do pai, mãe e filho, com isso o software carregou as informações dos marcadores SNPs do trio selecionado, posteriormente realizou a exportação dos dados de genotipagem para arquivos com extensões .txt, em seguida o sistema abriu a opção de seleção dos progenitores e prole que ao serem marcados foi possível escolher a forma de organização do resultado. Foram apresentadas em dois formatos: Separados por cromossomos ou separados por arquivos do tipo CHP[3], este formato foi a seleção escolhida. Próximo módulo está relacionado com a construções de nosso pipeline que realizou a detecção de mutações *de novo* a partir dos dados de genotipagem do SNP. Primeiramente, nosso algoritmo validou o *layout* do arquivo CytoScan HD™ (.CYCHP) em formato texto, posteriormente o algoritmo identifica as variáveis por meio das posições que estão disposto no arquivo, a partir da atribuições destas variáveis o algoritmo inicia as inferências sobre os achados de mutações *de novo* para cada marcador, a posteriori, são projetadas essas inferências para definir os tipos de substituições, além de indicar a origem do progenitor que gerou a mutação na prole. Este módulo finaliza com o carregamento das informações para uma base de dados. Para o fechamento deste estudo, o módulo se encerra com a análise de dados, por meio da extração, quantificação, aplicações de técnicas estatísticas, tais como: Regressão Linear, Clusterização e PCA (do inglês, *Principal Component Analysis*). A figura 8 demonstramos a metodológica deste estudo científico.

2 Arquivos .CYCHP possui as análises do protocolo GeneChip CytoScan HD™ contendo principalmente as informações sobre cromossomos; resultados das sondas; e genotipagem (com base no algoritmo BRLMM-P-Plus)

3 Aquivos .CHP contém os resultados (descrições dos algoritmos; sondas armazenadas na matriz; valores e informações do QC; nome do arquivo .CEL gerado e outros) das análises do conjunto de sondas gerados pelo ChAS

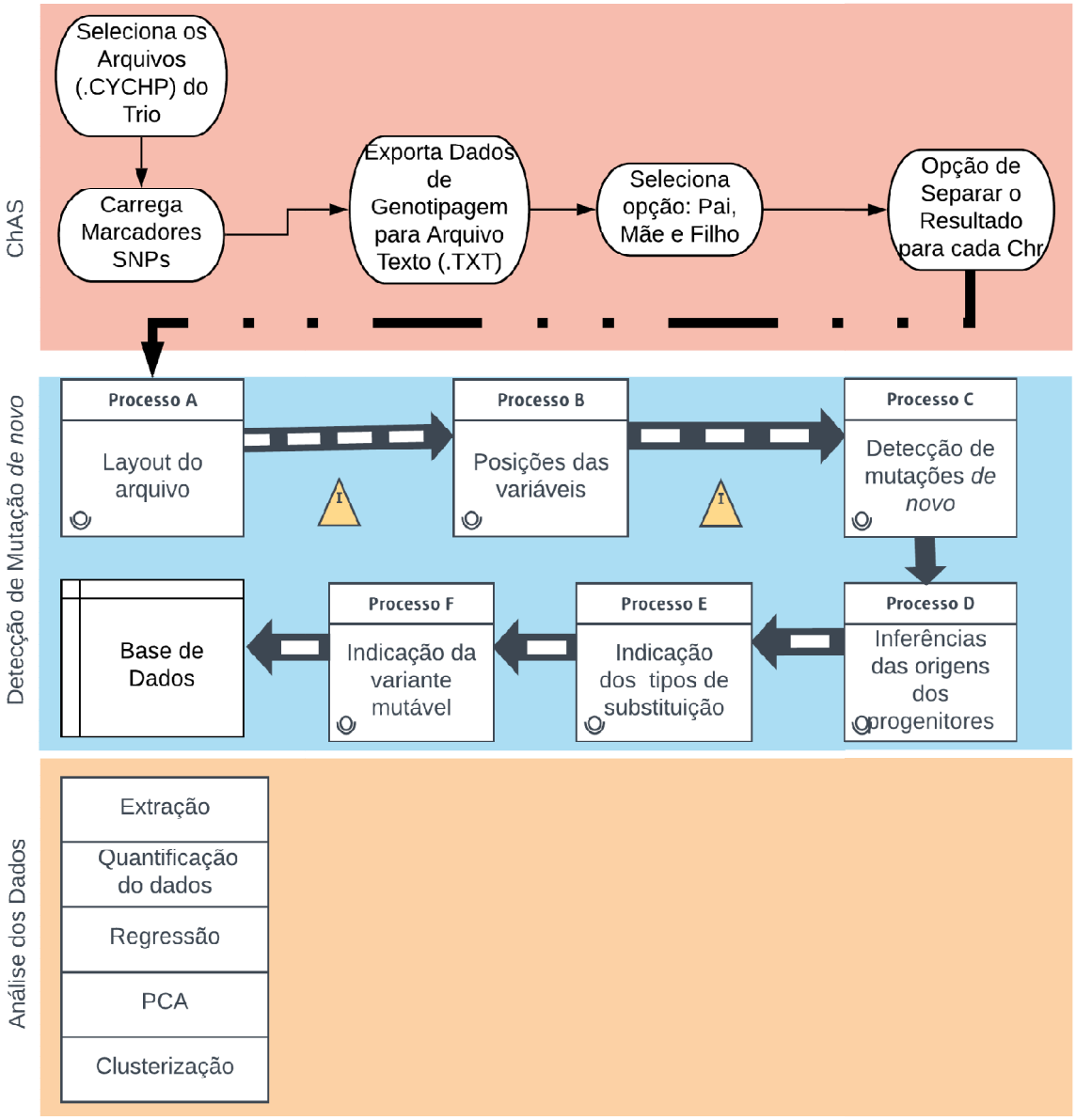


Figura 8. Fluxograma da organização metodológica realizada no presente estudo dividido em: *Workflow* do Software ChAS®; etapas para imputações das mutações *de novo*; fase de análise de dados para montagem de resultados.

4.2 Caracterização do Grupo Amostral

O grupo amostral da presente pesquisa, caracterizado como grupo caso, foi composto por 11 famílias, dos quais pelo menos um dos progenitores foi diretamente exposto à radiação ionizante durante o acidente radiológico do Césio-137, totalizando 37 indivíduos (11 pais, 11 mães e 15 filhos nascidos após o acidente). A dose absorvida de radiação ionizante para os indivíduos expostos variou de 0,2 a 0,5 Gy.

Um outro grupo de indivíduos não-expostos à radiação ionizante foi usado como controle. Esse grupo foi composto por 15 famílias goianas - compostas de 15 pais, 15 mães e 15 filhos – sem histórico de exposição à radiação ionizante.

O tamanho total dos indivíduos que participaram do presente estudo foi de 82 indivíduos.

4.3 Considerações Éticas

Os indivíduos separados em grupo caso e controle participaram voluntariamente do estudo. O presente estudo foi aprovado pelo Comitê de Ética e Pesquisa da Pontifícia Universidade Católica de Goiás apresentando o número do CAAE foi 49338615.2.0000.0037.

Cada amostra biológica, foi refrigerada a -20ºC, sendo que o material restante foi armazenado para estudos futuros nos termos da Resolução CNS Nº 441/11.

No momento da coleta, os participantes responderam voluntariamente a um questionário e assinaram a um Termo de Consentimento Livre e Esclarecido e/ou um Termo de Assentimento, quando pertinente.

4.4 Obtenção dos Dados

Neste estudo os arquivos. CEL contendo as informações dos genótipos dos indivíduos dos grupos exposto e controle foram disponibilizadas pelo NPR da PUC-GOIÁS. Sendo que a metodologia empregada para a obtenção dos arquivos. CEL é referente a plataforma GeneChip® CytoScan HD™ (Thermo Fisher Scientific, Massachusetts, EUA), seguindo as recomendações do fabricante.

Importante realçar que a amostra biológica usada para a obtenção do DNA genômico foi sangue periférico, colhido por função venosa a 10mL, tendo sua fração plasma, hemácias e anel leucocitário separados por centrifugação (12.000 rpm por 1 minuto).

O DNA foi extraído do anel leucocitário usando o kit de extração e purificação de DNA Illustra Blood GenomicPrep Mini Spin® (GE Healthcare Life Sciences, EUA). A quantificação da concentração de DNA genômico foi realizada em um espectrofotômetro

NanoVue® Plus (GE Healthcare, Life Sciences, Reino Unido). Ambos os procedimentos foram executados de acordo com os protocolos sugeridos pelos fabricantes.

4.5 Análise Cromossômicas por Microarranjos

A análise cromossômica por microarranjo (CMA) foi realizada usando-se o GeneChip® CytoScan HD™ (Thermo Fisher Scientific, Massachusetts, EUA), sendo que, a CMA executada no Software ChAS® em busca de alterações capazes de investigar alterações estruturais ao longo do genoma. Para isso, foram fixados como filtros: 15 e 8 marcadores de SNP's para se detectar microduplicações e microdeleções, respectivamente, distribuídos com uma média ≤2.000 pb, limitando-se a fragmentos ≥1kb.

A partir do arquivo .DAT, contendo o sinal bruto, foi gerado um arquivo .CEL, que contém as intensidades de sinal único para cada marcador, contendo as chamadas ao SNP foram lidos diretamente no software, ChAS®. Para a identificação de recursos e extração de sinal, os cartuchos de microarranjos GeneChip® foram digitalizados no equipamento Scanner GeneChip® e processados pelo pacote de software GeneChip® Command Console® (AGCC®, Affymetrix, USA).

Para o controle de qualidade (QC) foi aplicada as métricas usanda no microarranjo CytoScan®, recomendada pelo fabricante. Com isso, foi aplicado a métrica MAPD (do inglês, *Median of the Absolute values of all Pairwise Differences*), que representa o valor global da variação de todas as sondas dentro do microarranjos no genoma. Foi atribuído para a variável MAPD o valor menor ou igual a 0,25 para o CytoScan HD™ 750 K. Para a variável *Waviness* SD, que corresponde a uma medida global de variação de sondas dentro do microarranjo, sendo este insensível as variações e se concentra nas variações de longo alcance, o valor estipulado foi maior que 0,12. A última métrica do QC corresponde ao SNPQC, que quantifica a qualidade dos genótipos alelos que estão distribuídos no microarranjo, sendo que o valor para o controle desta variável foi maior ou igual a 15 para CytoScan HD™ 750 K.

Protoloco da geração dos resultados a partir da ferramenta ChAS® estão descritas no ANEXO I.

4.6 Estimativa dos Desvios Mendelianos

A estimativa dos desvios mendelianos (DM) foi possível extrair informações relevantes do trio (pai, mãe e filho(a)). Foi possível também exibir os dados dos

marcadores para um único cromossomo. Estes resultados foram transformados em informações que pudessem ser analisadas e tratadas de forma que as variações observadas nos SNPs pudessem ser corretamente identificadas. No presente estudo, DM foram estimados apenas para os autossomos.

Após a descoberta dos DM os algoritmos foram desenvolvidos na linguagem de programação PERL, para inferir a origem da mutação. Na Figura 9 abaixo, tempos a representação do fluxograma que descreve as etapas gerados pelos algoritmos para a montagem das estimativas dos DM.

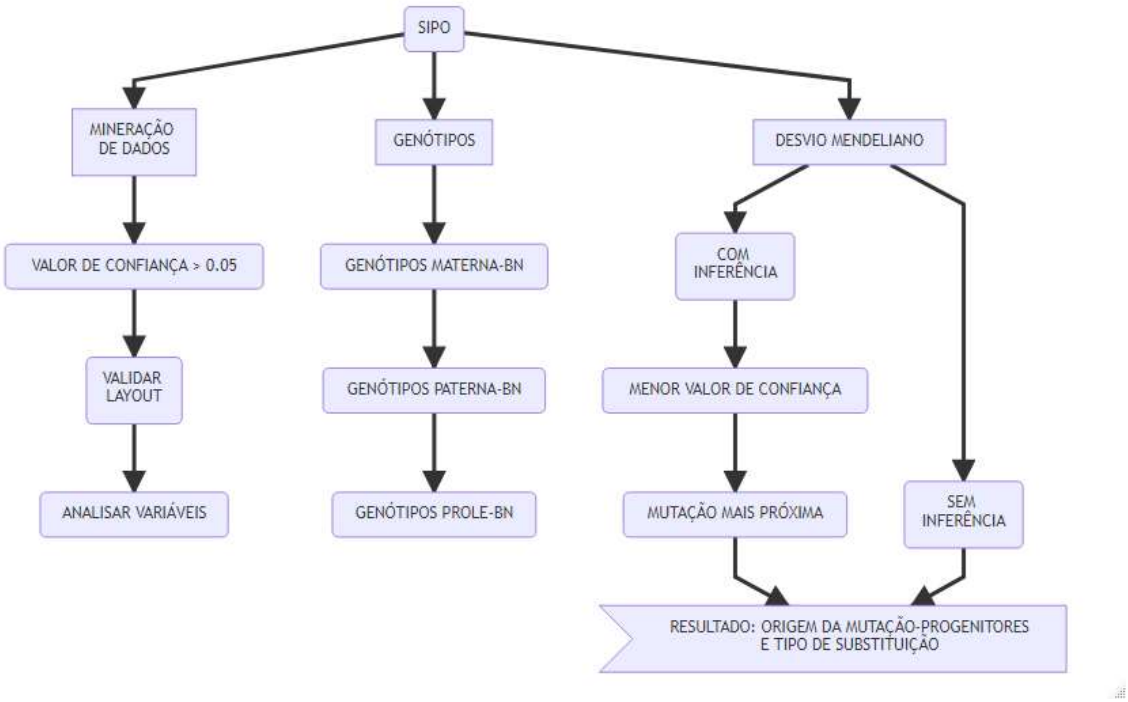


Figura 9. Fluxo das etapas geradas pelos algoritmos para identificação dos desvios mendelianos por SNP.

Aplicamos a solução recursiva, pois o tamanho de entrada é considerada grande por termos números de possibilidades entre as variáveis consideradas (genótipos maternos e paternos são 2^{10} combinações para cada um), o que torna relevante a ordem de crescimento do tempo de execução do SIPO. Desta forma, comparamos os genótipos maternos (2^{10} combinações) X genótipos paternos (2^{10} combinações) e comparamos com os genótipos das proles (2^9 combinações quando os progenitorese são homozigoses e 2^7 quando os progenitores são heterozigoses).

Com isso foi possível inferir a mutação realizando a classificação do tipo da substituição das bases nitrogenadas geradas no SNP e sua origem parental. No Quadro 1 observamos os possíveis cenários para identificar os desvios mendelianos.

Quadro 1 Combinações que foram geradas pelo scritp nas mutações de bases nitrogenadas

Mutação Base Nitrogenada - Origem Materna ou Paterna	Para este caso, não se consegue determinar a origem da mutação. Pois, os valores de confiança dos progenitores eram iguais, além, da análise da menor distância euclidiana entre as posições não terem sido conclusivo. Ex: Paterna (AA)/Materna (AA)/Herdeiro (AC)
Mutação Base Nitrogenada - Origem Materna	Para este caso, mostra a evidência que a origem da mutação é materna. Ex: Paterna (AC)/Materna (AA)/Herdeiro (CC)
Mutação Base Nitrogenada - Origem Paterna	Para este caso, mostra a evidência que a origem da mutação é materna. Ex: Paterna (AA)/Materna (CC)/Herdeiro (GC)
Transição - Purina/Purina	A mutação ocorre entre Purinas (A/G) ou (G/A)
Transição - Pirimidina/Pirimidina	A mutação ocorre entre Pirimidina (C/T) ou (T/C)
Transversão - Purina/Pirimidina	A mutação ocorre entre Purinas e Pirimidinas (A/C ou G/T), e vice-versa

Utilizando o mesmo resultado, foi aplicado outro algoritmo, baseado também em lógica booleana e soluções recursivas, aplicado para a informação de genótipo (*genotype call: AA|AB|BB)* para trio. Para esta situação, tivemos a identificação de mutações de genotipagem, informando os desvios mendelianos, além da origem da mutação (paterna ou materna). No Quadro 2 foram esboçados as possíveis combinações geradas por este algoritmo.

Quadro 2 Combinações geradas pelo *scritp* nas mutações de genotipagem

Mutação Genotípica - Origem Materna ou Paterna	Para este caso, não se consegue determinar a origem da mutação, pois analisando os valores de confiança e a menor distância euclidiana, ambos os casos, são inconclusivos. Ex: Paterna (AA)/Materna (AA)/Herdeiro (AB) ou (BB)
Mutação Genotípica - Origem Materna	Para este caso, mostra a evidência que a origem da mutação é materna. Ex: Paterna (AA)/Materna (BB)/ Herdeiro (AA)
Mutação Genotípica - Origem Paterna	Para este caso, mostra a evidência que a origem da mutação é Paterna. Ex: Paterna (BB)/Materna (AA)/ Herdeiro (AA)

O fluxograma com a representação dos procedimentos que ocorrem nos algoritmos para a detecção do DM deste estudo científico, está representado na Figura 10.

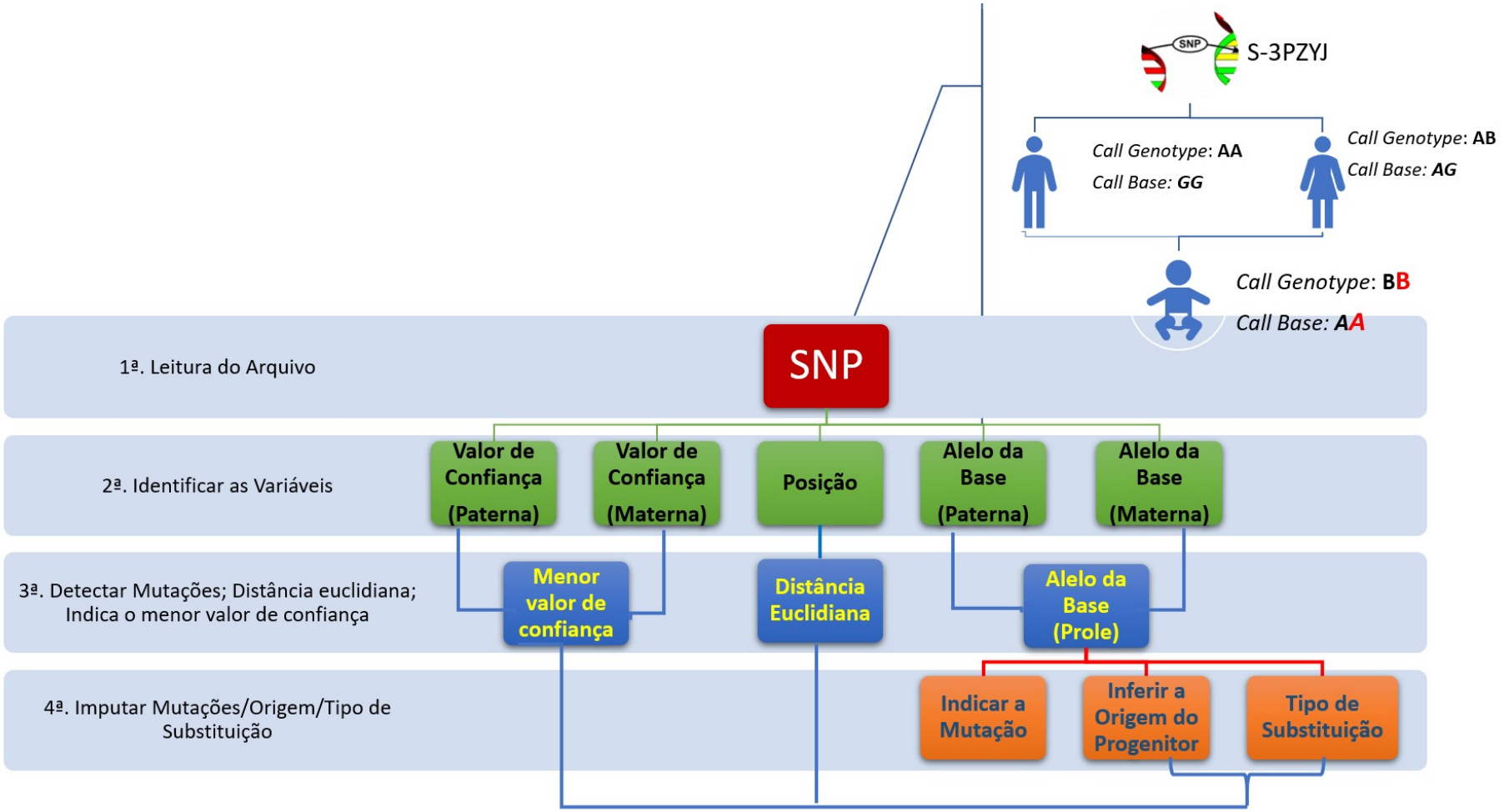


Figura 10. Fluxo das etapas que demonstram as etapas dos algoritmos para os achados de desvios mendelianos para cada trio.

Posteriormente, carregamos todos os resultados em um base de dados (MySql©), com as informações das mutações dos trios para geração de estatísticas e relatórios conclusivos. Assim, demonstramos o uso de gráficos exploratórios, em formas de gráficos de caixas, gráficos de densidade, probabilidades, validação de testes estatísticos usando a programação R©. O fluxograma com a descrição do funcionamento dos procedimentos adotados para execução deste estudo científico, está representado na Figura 11.

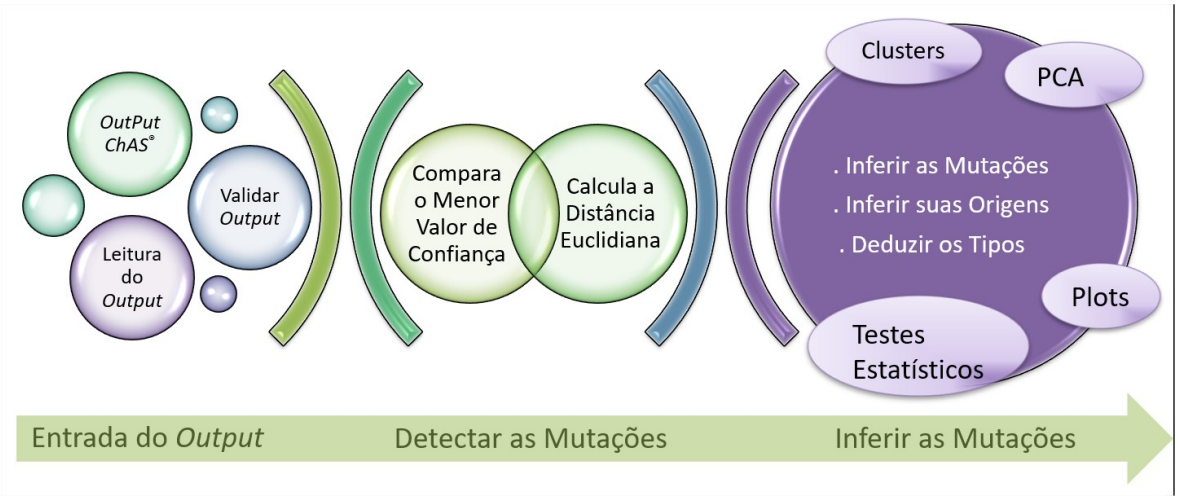


Figura 11. Fluxo dos procedimentos adotados para a execução dos *pipelines* para detecção de erros mendelianos.

Frequência Média de Desvio Mendeliano

A frequência média do desvio mendeliano (FM_{DM}) *de novo* representa a relação do somatório de desvio mendeliano por SNP com o produto do *lócus* bialélicos e o total de SNPs válidos. No presente estudo foram incluídas somente desvio mendeliano *de novo*, correspondendo às mutações germinativas nos autossomos da prole nascida de progenitores expostos acidentalmente à radiação ionizante de césio-137. A FM_{DM} conforme representada na equação 3: (COSTA *et al.,* 2011; DA CRUZ *et al.,* 2008).

Equação 3

$$FM_{DM} = \frac{\sum T_{DM}}{bxnsv}$$

ΣT_{DM} : Total de desvio mendeliano por geração.

b: lócus bialélica (2)

nsv: número de sondas válidas[4] para o microarranjo de acordo com a montagem da sequência humana de referência (GRCh37/hg19).

A Equação 3 foi aplicada para os achados de desvios mendelianos. Para se avaliar o impacto acumulado dos ganhos e perdas genômicos, foi analisado o *burden*[5] dos desvios mendelianos que correspondeu o total de desvios mendelianos (FM_{DM}), somando-se os números de desvios mendelianos das proles.

Prunning de SNP's Baseado no Desequilíbrio de Ligação

Geramos os subconjuntos com *prunning* de SNP's que apresentaram equilíbrio aproximado de ligação entre si, ou seja, considerou as correlações e combinações lineares entre SNP's. Primeiramente aplicamos o *pipeline* do pacote PLINK (APÊNDICE II) até utilizar o desbaste com os seguintes parâmetros para DL máximo de 0,1, usando 500 SNP's de janelas deslizantes e incremento de 5 a cada etapa.

4.9 Análise estatística

O teste paramétrico de Shapiro-Wilk foi utilizado para determinar que o conjunto de dados analisados neste estudo, dada as variáveis aleatórias, foi modelado por uma distribuição normal. O teste paramétrico de *Student-t* foi aplicado para comparar as médias de dois grupos (Caso e Controle). O conjunto da FM_{DM} do presente estudou e suas médias que foram representadas conforme uma distribuição normal. No presente estudo, foi utilizado o teste *F* para comparar as variações das duas amostras de populações aplicadas à uma distribuição normal.

A regressão linear foi usada para analisar a relação entre preditores (idade dos pais à época da concepção da prole) de escala de intervalo e resultados da FM_{DM} (frequência média de desvios mendelianos). Para esta técnica, o conjunto de dados do presente estudo não apresentou valor significativo para caso ($p \cong 0{,}249$) e para controle ($p \cong 0{,}248$). Portanto, o método de mínimo quadrado foi aplicado para encontrar um melhor ajuste para este conjunto de dados, tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados.

Todas as análises foram realizadas utilizando o pacote estatístico R, com nível de significância de 5% ($p<0{,}05$).

4 Sondas válidas são marcadores SNPs em que o valor de confiança para cada marcador foi menor que $5x10^{-2}$

5 *burden* é definido por um número *de novo* na substituição da base em um SNP nas proles em que os progenitores foram expostos pela radiação ionizante.

5. RESULTADOS

As médias das idades dos progenitores à época da concepção foram 31,4 e 32,2 anos para os pais e 26,4 e 27,5 anos para as mães dos grupos caso e controle, respectivamente. As doses individuais absorvidas para o grupo exposto variaram de 0,2 a 0,5Gy. No grupo controle foram incluídas 15 famílias (trios), correspondendo a 45 indivíduos, relativos aos progenitores e a uma criança. Os participantes do grupo controle não apresentavam história de exposição acidental, ocupacional ou para fins terapêuticos ou de diagnóstico à radiação ionizante. Os casos e controles eram residentes da cidade de Goiânia-Goiás. As médias da idade da geração F1 foram de 14,2 e 10,5 para casos e controles, respectivamente. O quadro 3 e tabela 2 contém os dados gerais e descritivos dos grupos participantes.

Quadro 3 Dados gerais dos grupos caso e controle para as gerações parental e F1 incluídos no estudo da sobre a indução de mutação germinativa na prole de indivíduos expostos acidentalmente a doses baixas de radiação ionizante de césio-137.

Geração	Variáveis		Casos	Controle
Parental	n		11	15
	Intervalo etário (anos)		16 a 56	19 a 55
	Média das idades (anos) à concepção (±DP)	Paterna	31,4 (12,3)	32,2 (12,8)
		Materna	26,4 (5,1)	27,5 (9,8)
	Dose absorvida (Gy)		0,2 a 0,5	0
F1	n		15	15
	Intervalo etário (anos)		2 a 20	0,85 a 26
	Média das idades (anos) (±DP)		14,1 (6,1)	10,5 (8,5)
	Proporção entre os sexos (H/M)		8/7	10/5
	Média da FM_{DM} (±DP)		1.3×10^{-3} ($\pm 0.4 \times 10^{-3}$)	0.9×10^{-3} ($\pm 0.2 \times 10^{-3}$)

DP: Desvio Padrão; H: Homens; M: Mulheres;

Tabela 2. Dados gerais dos grupos controle e exposto a respeito do estudo de mutação da linha germinativa em filhos de pessoas acidentalmente expostas a baixas doses absorvidas de radiação ionizante de césio-137 em Goiânia (Brasil).

Grupo	Família	Progenitor Exposto	Dose Absorção (Gy)	Idade Paterna[2,*]	Idade Materna[2]	Idade da Prole	Sexo da Prole	DMs[1]				Total de SNPs Válidos	Frequência de DMs
								Paterna	Materna	Desconhecido	Total		
Controle	Ct001	Nenhum	0	40	36	9	Feminino	783	694	10	1487	702,304	1,06E-03
	Ct25	Nenhum	0	47	36	9	Masculino	545	631	3	1179	683,381	8,62E-04
	Ct27	Nenhum	0	26	26	23	Masculino	594	729	9	1332	692,311	9,62E-04
	Ct39	Nenhum	0	24	24	3	Feminino	679	679	11	1369	674,320	1,02E-03
	Ct40	Nenhum	0	45	37	3	Masculino	710	711	30	1451	696,544	1,04E-03
	Ct45	Nenhum	0	37	31	15	Masculino	643	663	8	1314	683,243	9,62E-04
	Ct51	Nenhum	0	35	34	2	Feminino	384	479	8	871	697,737	6,24E-04
	Ct52	Nenhum	0	55	41	1	Masculino	1015	588	32	1635	710,477	1,15E-03
	Ct53	Nenhum	0	35	27	8	Masculino	315	361	6	682	713,372	4,78E-04
	Ct60	Nenhum	0	40	38	1	Feminino	501	482	6	989	712,261	6,94E-04
	Ct66	Nenhum	0	31	20	26	Feminino	543	594	2	1139	707,798	8,04E-04
	Ct68	Nenhum	0	33	20	10	Masculino	758	654	11	1423	712,191	1,00E-03
	Ct70	Nenhum	0	20	24	8	Feminino	448	545	3	996	714,892	6,96E-04
	Ct72	Nenhum	0	31	20	14	Feminino	521	614	4	1139	710,259	8,02E-04
	CtF09	Nenhum	0	19	21	25	Masculino	718	696	9	1423	712,352	9,98E-04
Expostos	Ex04	Paterna	0.1	27	27	20	Masculino	900	966	7	1873	698,305	1,34E-03
	Ex06	Paterna	0.3	35	26	9	Masculino	1045	1082	9	2136	693,858	1,54E-03
	Ex07-1F	Materna	0.2	54	24	19	Masculino	976	850	8	1834	692,375	1,32E-03
	Ex07-4F	Materna	0.2	56	26	17	Feminino	1509	1085	18	2612	689,467	1,89E-03
	Ex08	Materna	0.2	18	20	8	Masculino	538	718	4	1260	706,759	8,92E-04
	Ex10	Materna	0.2	21	24	2	Feminino	1187	1054	21	2262	705,993	1,60E-03
	Ex12	Materna	0.3	31	30	3	Masculino	1361	1486	28	2875	693,463	2,08E-03
	Ex15	Paterna	0.2	18	27	16	Masculino	560	645	7	1212	706,780	8,58E-04
	Ex18	Paterna	0.2	47	30	18	Masculino	819	800	12	1631	707,366	1,15E-03
	Ex21	Materna	0.2	38	27	20	Feminino	457	513	2	972	707,827	6,86E-04
	Ex22-2F	Materna	0.2	29	31	20	Feminino	1006	664	3	1673	707,075	1,18E-03

	Ex22-3F	Materna	0.2	32	34	17	Feminino	845	566	3	1414	708,278	9,98E-04
	Ex22-4F	Materna	0.2	33	35	16	Masculino	781	518	5	1304	708,885	9,20E-04
	Ex24	Paterna	0.5	21	19	12	Feminino	1010	1102	47	2159	703,635	1,53E-03
	Ex25	Paterna	0.5	18	16	15	Feminino	598	708	10	1316	706,598	9,32E-04

[1]Desvio Mendeliano; [2]Idade da Concepção; *Todas as idades estão em anos.

92

O presente estudo investigou os números de desvios mendelianos (DM) associados aos autossomos dos grupos amostrais. Portanto, no estudo foram investigados os 22 cromossomos. Os cromossomos sexuais foram descartados da análise, pois o X apresenta bastante ruído em seus dados de SNP e o Y por ter uma baixa cobertura de marcadores. Neste contexto, o *burden* da frequência de mutação germinativa em DM ($FM_{DM\beta}$) foi de aproximadamente 44% maior para a progênie dos indivíduos expostos em relação aos controles saudáveis da população de Goiânia.

Para testar a distribuição normal do conjunto de dados, foi aplicado o teste de Shapiro-Wilk sobre a variável número de desvios mendelianos no conjunto de dados, cujo *p-value* apresentado (p=0,1492 e w=0,9127) para grupo caso e (p=0,2773 e w=0,9304) para o grupo controle, portanto, maior que $p < 0,05$, rejeitando-se a hipótese nula. Desta forma, pode-se concluir que os dados estavam distribuídos de forma normal. Abaixo, na Figura 12, aplicaremos o recurso do gráfico tipo QQ-Plot (*Quantile-Quantile*) para demonstração da distribuição normal dos dados.

a) b)

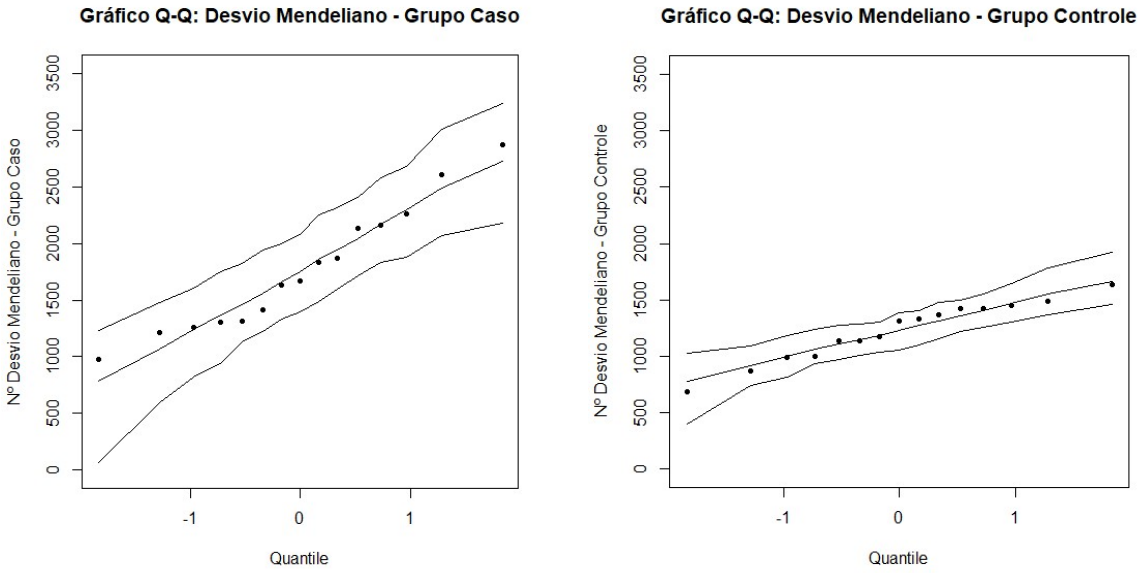


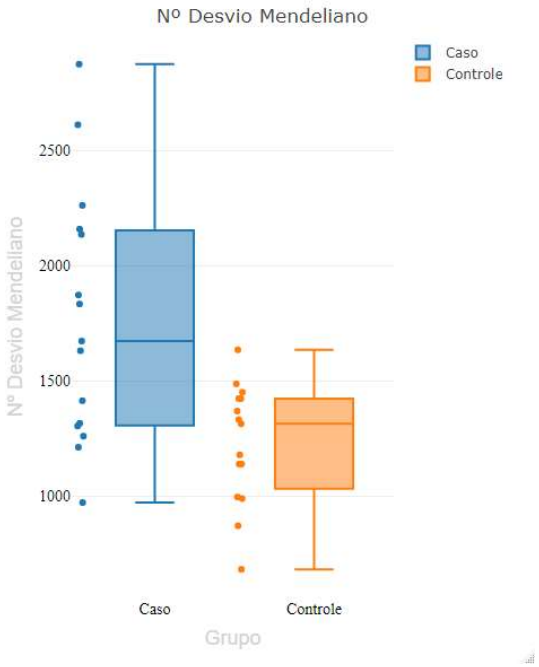
Figura 12. Gráfico QQ representando os valores dos quartis referente ao número de desvios mendelianos por grupos caso (a) e controle (b).

O teste paramétrico de *Student-t* foi usado para comparar as médias de dois grupos sob o pressuposto de que ambas as amostras são aleatórias, independentes e provêm de uma população normalmente distribuída com variações desconhecidas. A diferença na distribuição das frequências da FM_{DM} entre casos e controles mostrou-se estatisticamente

significativa para as diferenças entre as médias ($p<2x10^{-3}$) ao nível de significância/confiança de 95% de acerto.

A análise dos dados permitiu a identificação de 26.533 e 18.429 DM, para os grupos de casos e controles, respectivamente. Todas os DM observadas para os grupos de casos e controles do presente estudo foram incluídas nas análises estatísticas. Os menores números de DM encontrados foram 972 e 682, para os grupos casos e controles respectivamente. Contudo, os maiores números de DM foram 2.875 e 1.635, para os grupos casos e controles respectivamente (Figura 13a e 13b). O teste *F* para comparar as variações das amostras (caso e controle) com distribuição normal mostrou que as variações ($F=4,47$; $p<8x10^{-3}$) entre caso e controle foi significativamente diferente.

a)



b)

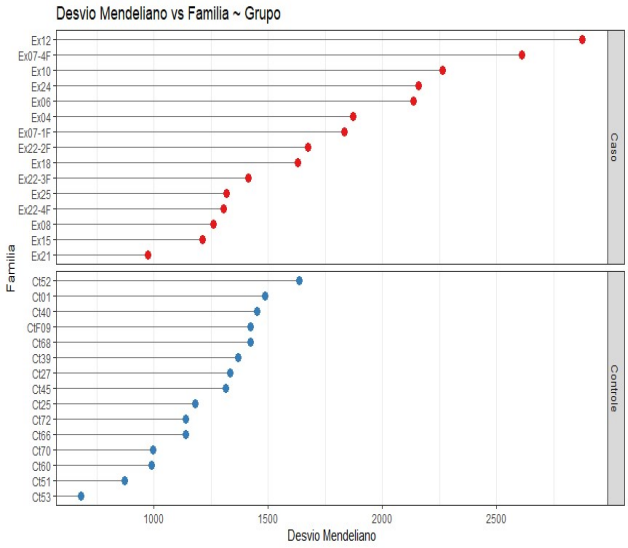


Figura 13. a) Médias dos números de desvios mendelianos da progênie de caso e controle, expostos às dosagens baixas de radiação ionizantes do Césio-137 em Goiânia (Brasil). b) Média das frequências de mutação germinativa por desvios mendelianos observados nos autossomos da progênie de casos expostos às doses baixas de radiação ionizante de césio-137 e controles de Goiânia (Brasil).

Também realizamos uma regressão linear para avaliar a relação entre as doses absorvidas de radiação e o FM_{DM} em nossas coortes. Nossos resultados foram estatisticamente significativos (p = 0,004; R^2 = 0,257), sugerindo que baixas doses

absorvidas de RI poderiam predizer um aumento do desvio mendeliano no grupo exposto, o qual poderia ser ajustado linearmente (Figura 14) seguindo a equação abaixo:

Equação 4

$$MF_{MD} = 0.001 + 0.001(dose)$$

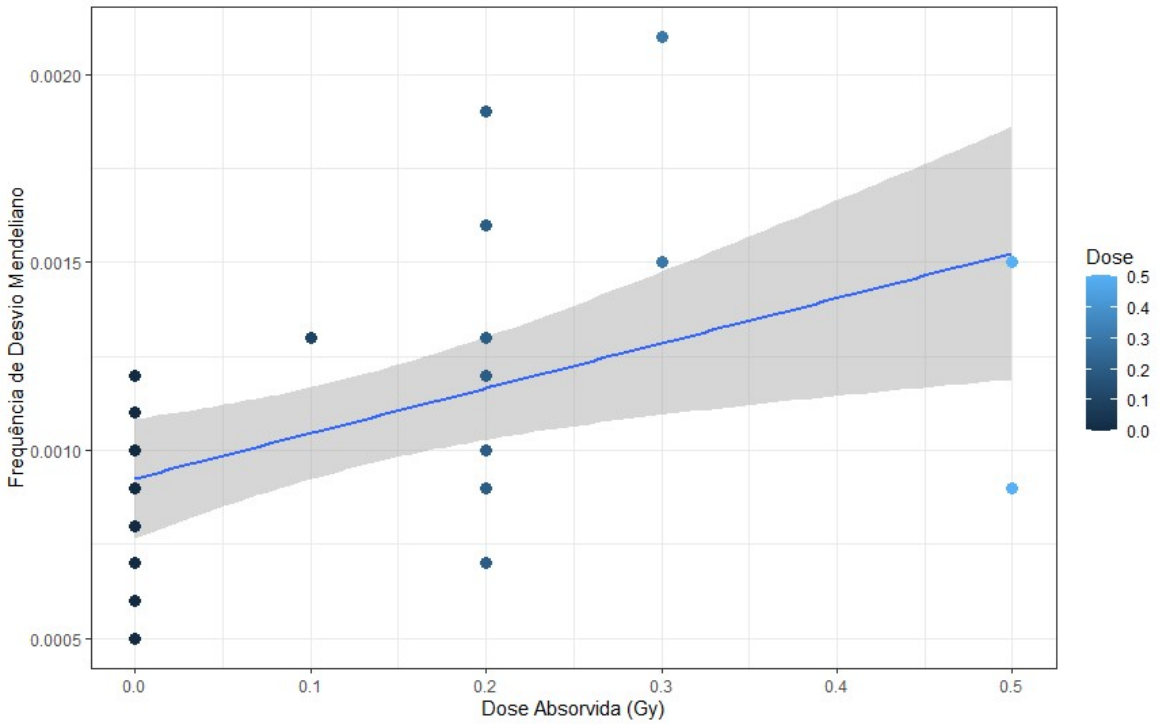


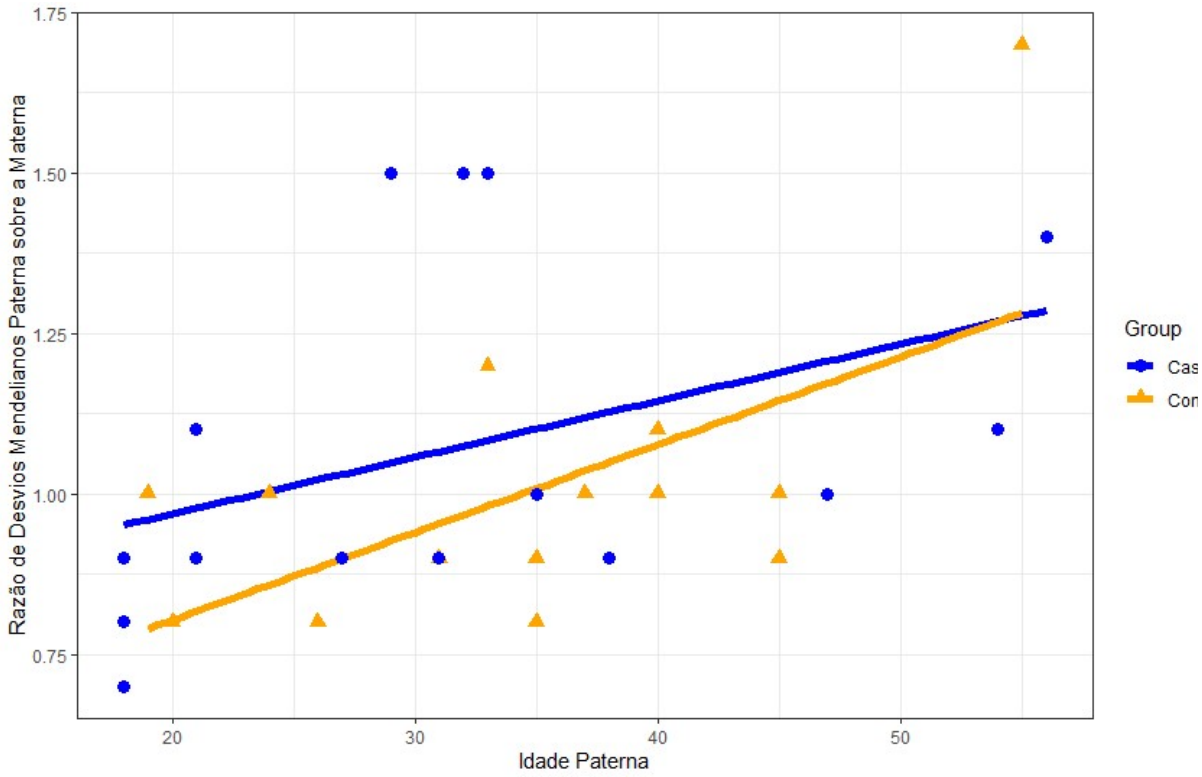
Figura 14. Representação da relação entre as doses absorvidas pela radiação e a frequência média dos desvios mendelianos em uma coorte de pessoas concebidas após a exposição dos pais à radiação ionizante.

Em nosso estudo, o sexo dos progenitores não teve efeito sobre o FM_{DM} de SNPs autossômicos, pois tanto para os grupos de caso quanto para os de controle, mães e pais contribuem com números iguais de DM *de novo* para seus filhos. Quando levado em consideração o sexo do pai exposto, a média das frequências de mutações da linha germinativa de crianças nascidas de pais expostos foi 1,2 $x10^{-3}$ (± 0,3 $x10^{-3}$) e para mães expostas foi 1,3 $x10^{-3}$ (± 0,5 $x10^{-3}$), sem diferenças estatísticas (p = 0,195) intragrupo.

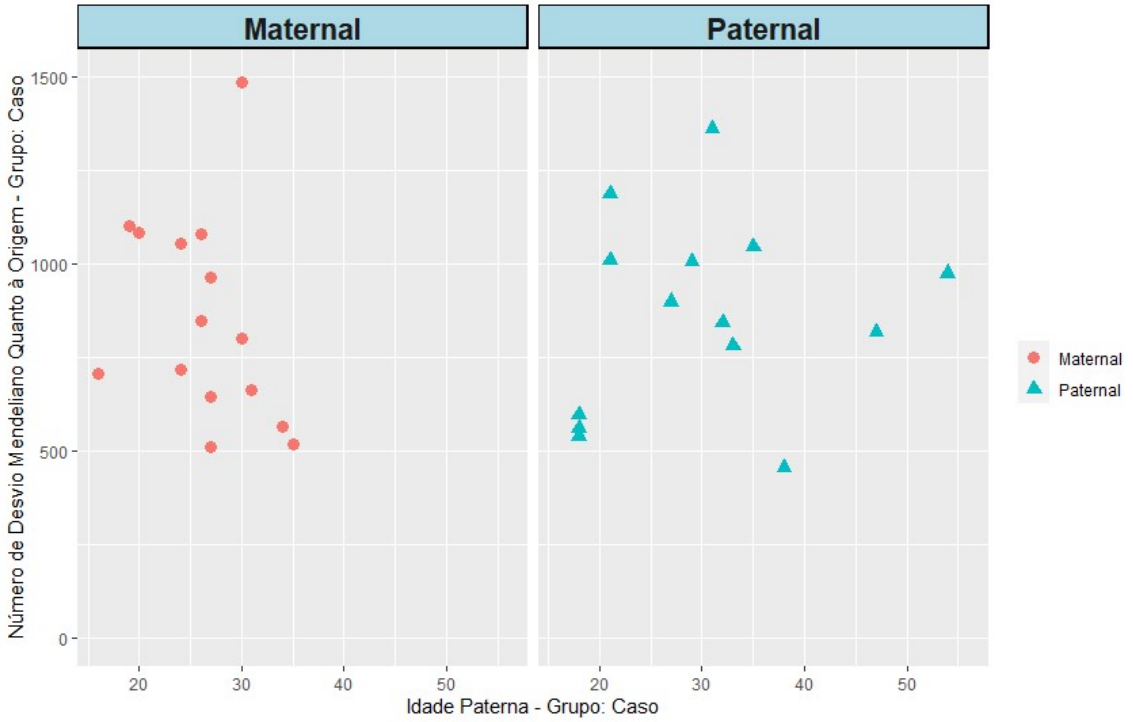
Com relação ao efeito potencial da idade dos pais, nosso grupo de controle revelou que pais mais velhos contribuíram com mais DMs para seus filhos (Figuras 15A-15C), o que poderia ser modelado pelo número de divisões de espermatogônias mitóticas em

função da idade, reforçando achados anteriores considerados viés de mutação masculina (CAMPBELL e EICHLER, 2013; JÓNSSON *et al*., 2017).

A)



B)


Maternal
Paternal
Número de Desvio Mendeliano Quanto à Origem - Grupo: Caso
1500
1000
500
0
20
30
40
50
Idade Paterna - Grupo: Caso
Maternal
Paternal

C)

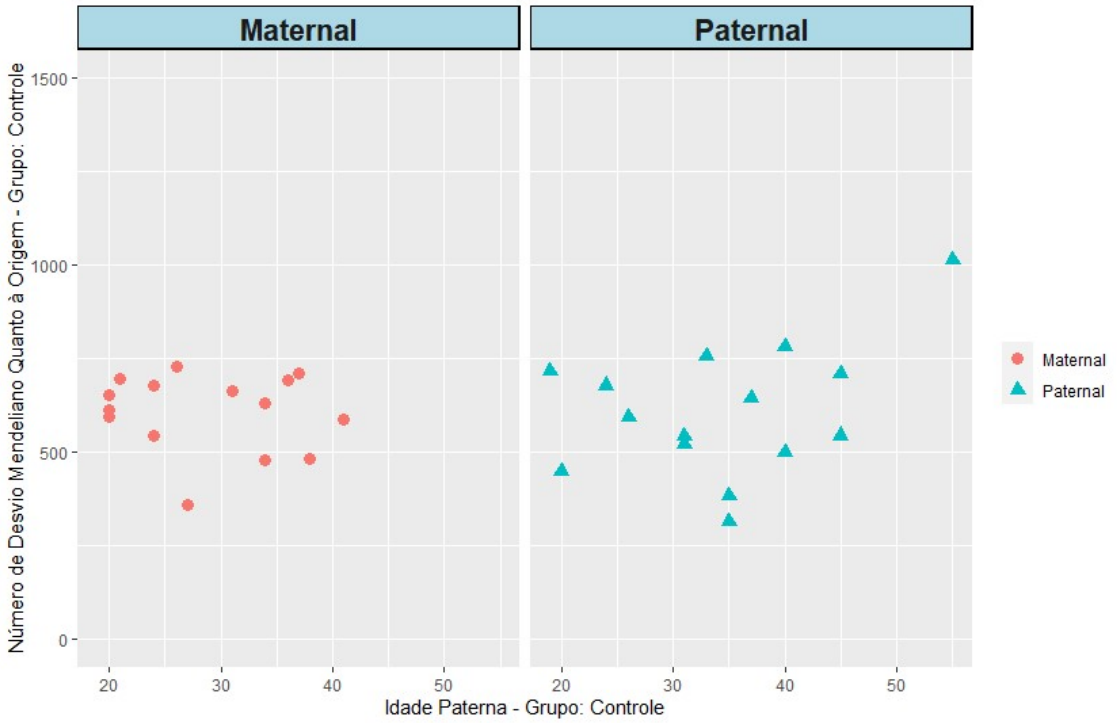

Maternal
Paternal
Número de Desvio Mendeliano Quanto à Origem - Grupo: Controle
1500
1000
500
0
20
30
40
50
Idade Paterna - Grupo: Controle
Maternal
Paternal

Figura 15. Efeito potencial da idade parental. (a) A razão de mutações paternas em função da idade paterna na concepção. Cada ponto representa os dados de uma criança (prole) com idades parentais semelhantes. A posição do eixo x é a idade dos pais, a posição do eixo y é o total do número de desvio mendeliano de origem paterna e materna. Demonstra a transmissão de dados sobre a origem parental que transmitiu a mutação à criança no grupo de caso (b) e no grupo de controle (c).

Na figura 16, representamos, por meio da regressão linear, o efeito das idades parentais sobre os achados de DMs *de novo*.

A) B)

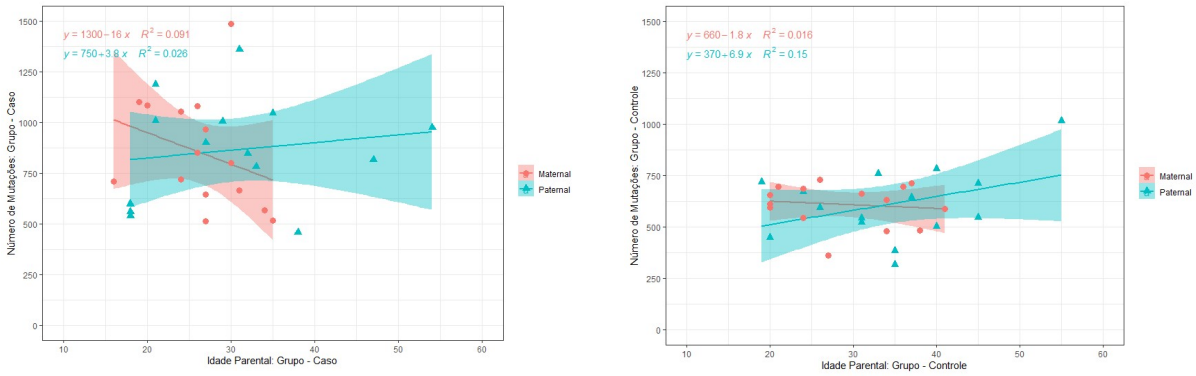


Figura 16. Representação do efeito da idade paterna e materna sobre o número de desvio mendeliano, em relação à origem parental, para o grupo caso (a) e controle (b).

Na figura 17 temos a representação das proporções de substituições nas bases nitrogenadas encontras nos DM do grupo caso (a) e grupo controle (b).

a) Caso b) Controle

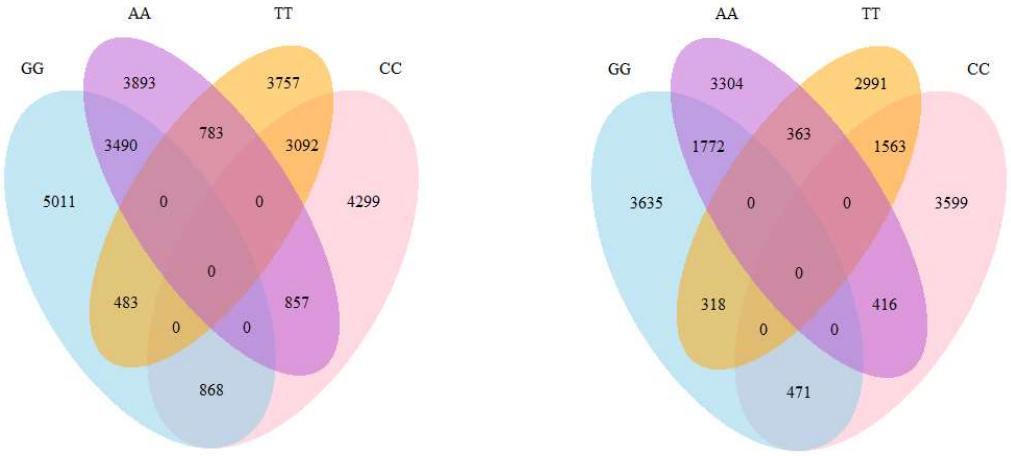


Figura 17. Temos a representação em forma de diagrama de Venn dos números de bases nitrogenadas a partir dos DM. A diferença entre os grupos ocorre somente na proporção de números.

Nesse contexto também, a FM_{DM} apresentou-se como um marcador sensível o suficiente para separar o grupo de crianças nascidas de progenitores expostos a RI de controles nascidos de progenitores não expostos da mesma população.

Montamos um dataset com 522.172 SNPs utilizando os recursos de filtros com qualidades, que incluiram o *call rate* > $5x10^{-2}$, eliminação de marcadores com informações nulos ou vazios e SNPs que apresentaram mutações em suas proles.

Este dataset foi utilizado para demonstrar que os grupos caso e controle são da mesma população. Para esta comprovação do efeito da qualidade das amostras, aplicamos os procedimentos baseados em desbaste de DL (do inglês, *Linkage Disequilibrium*) entre SNP. Os dados de DL indicaram que os grupos caso e controle pertenciam à mesma população (Figura 18).

a) b)

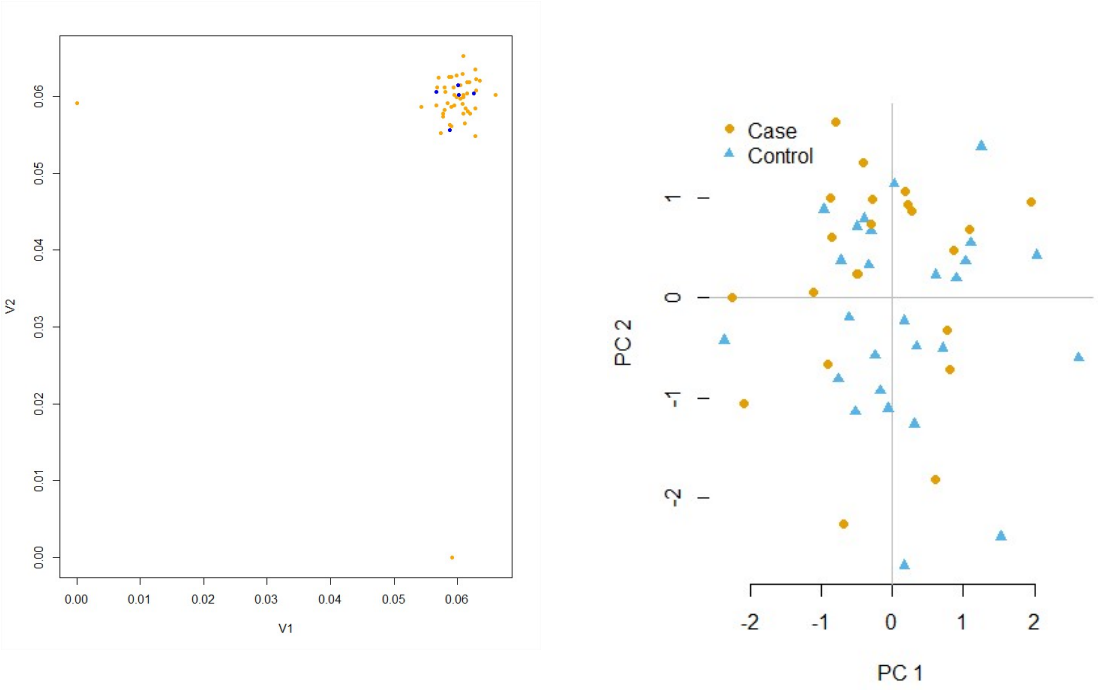


Figura 18. a) Cluster baseado na distância aos pares da identidade-por-estado (IBS). b) PCA de 2,7 K snps, resultado do desbaste usando a janela de 500 snps; com incrementos de 5 SNPs e LD máximo de 0.1. As variáveis contidas no PCA representa a matriz de relacionamento padronizada por variância; dimensionamento multidimensional (MDS) com base nas distâncias de *Hamming*.

Aplicamos a análise de variância para testar se houve existência de igualdade nas médias das idades paternas e maternas entre os grupos. Utilizamos o teste múltiplo de Tukey (Figuras 19A-19D), por realizar várias comparações em pares, para estimar as diferenças. O intervalo de confiança para o teste foi de 95%. O resultado gerado consiste

que a diferença entre as médias das idades dos progenitores entre os grupos não foi significativa (t = 1,238; p<2x10^{-1}).

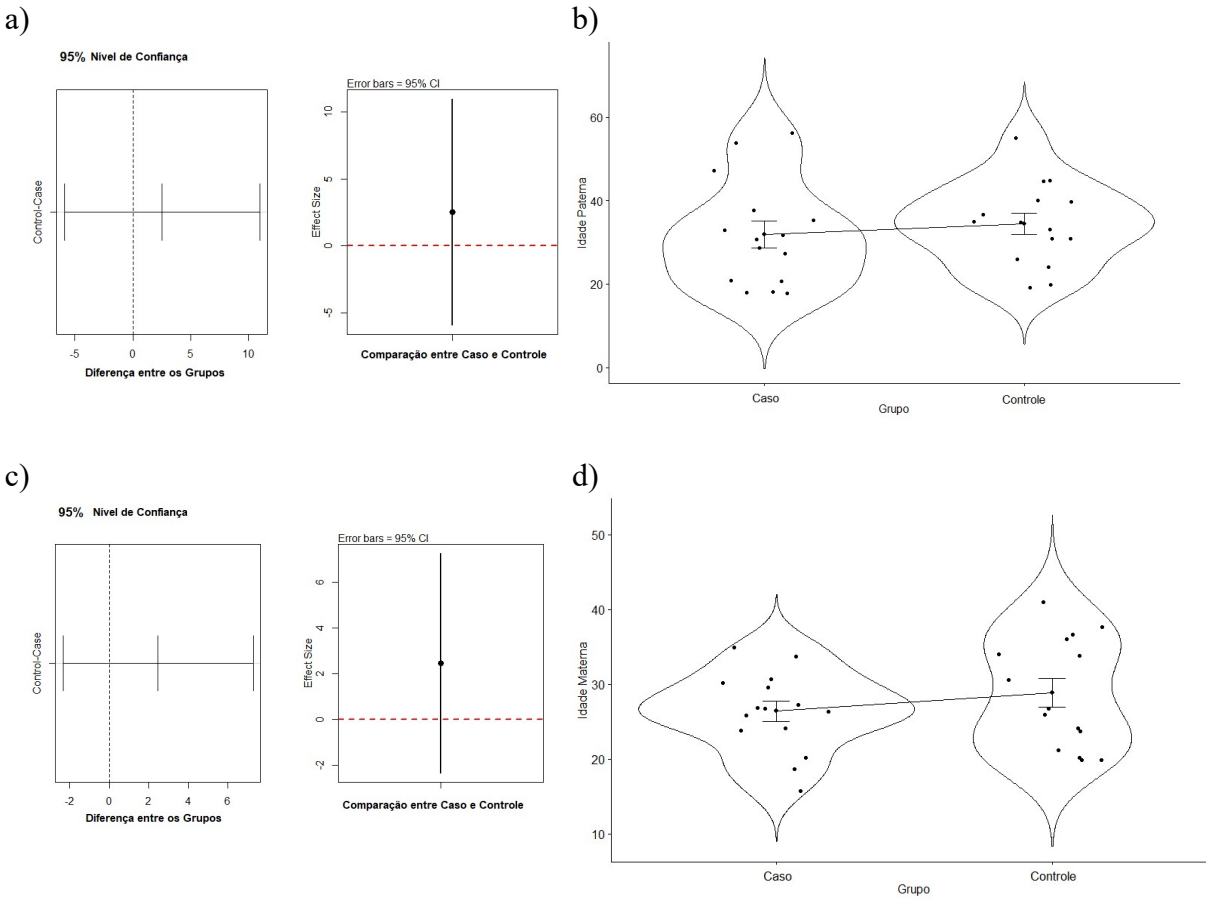


Figura 19. Os gráficos representam a técnica de Tukey para realizar múltiplas comparações usando as variáveis idade paterna (a) e idade maternal (c) e grupo. Neste caso, os valores caso e controle não apresentaram divergências em suas médias. Representamos a densidade de Kernel analisando as probabilidades das observações encontradas nas amostras para a idade paterna (b) e para a idade maternal (d), destacamos que não há diferença significativa entre as medias ($p \cong 0{,}54$ para idade paterna e $p \cong 0{,}3$ para a idade materna).

Trabalhos publicados sobre tipos de substituições em bases nitrogenadas de DNA estão bem representadas na linhagem germinativa e ocorrem espontaneamente (BROVARETS & HOVORUN, 2015). Estudos anteriores sugerem que as taxas de substituição de transições tendem a ser maiores (DA CRUZ, 1997) do que o esperado por acaso em relação às transversões (LYONS & LAURING, 2017). Na Figura 20, temos os resultados referente aos achados, que reforçam esta observação, pois foi possível identificar uma maior proporção de transições (74,5%) para os casos em relação aos controles (25,5%).

a) b)

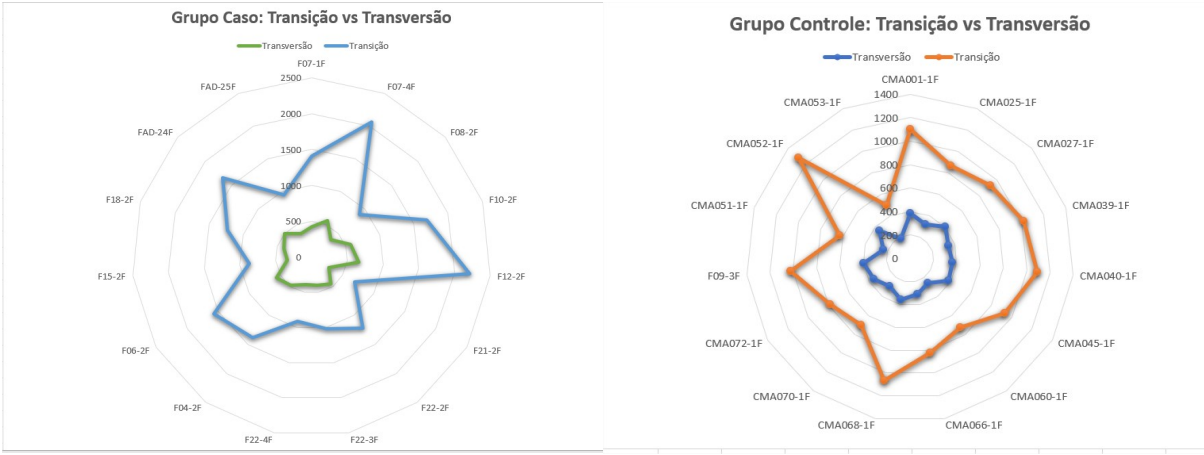


Figura 20. Os resultados representam a proporção entre os tipos de substituições: Transição e Transversão, respectivamente do grupo caso (a) e grupo controle (b).

Os achados do presente estudo corroboram essas observações, uma vez que uma proporção maior de transições foi observada nas crianças de casos e controles.

Em geral, presume-se que, em grupos de amostras pequenas, seria muito difícil detectar o efeito da idade materna sobre a carga de mutações pontuais na prole. No entanto, a fim de testar a hipótese de que em nossa coorte exposta mutações da linha germinativa em ambos os sexos foram induzidas por danos pela exposição a baixas doses de IR, estratificamos nosso conjunto de mutações de novo em 6 classes com base em alelos parentais e derivados (Tabela 3).

Tabela 3. Resumo dos dados descritivos dos grupos caso e controle para as seis classes de substituição de bases no genoma de crianças concebidas após exposição dos pais a baixas doses de radiação ionizante e seus controles.

Grupo	Classe	Mínimo	Máximo	Média	Desvio Padrão	Total
Controle	C>A	38	137	100,93	29,058	1.514
	C>G	45	141	111,47	28,538	3.344
	C>T	238	558	435,93	92,669	2.798
	T>A	45	84	64,67	14,034	6.539
	T>C	252	635	418,93	99,427	970
	T>G	61	141	96,67	21,091	6.284
Exposto	C>A	72	209	148,27	45,325	2.224
	C>G	90	291	168,60	56,616	2.529
	C>T	376	1.166	672,67	234,248	10.090
	T>A	48	147	85,27	27,825	1.279
	T>C	307	847	561,60	161,834	8.424
	T>G	75	219	132,47	42,797	1.987

Todas as transições e transversões tenham sido observadas em nosso conjunto de dados (Figura 21), sendo que, C> T e T> C foram super-representados, tanto para os casos quanto para os controles.

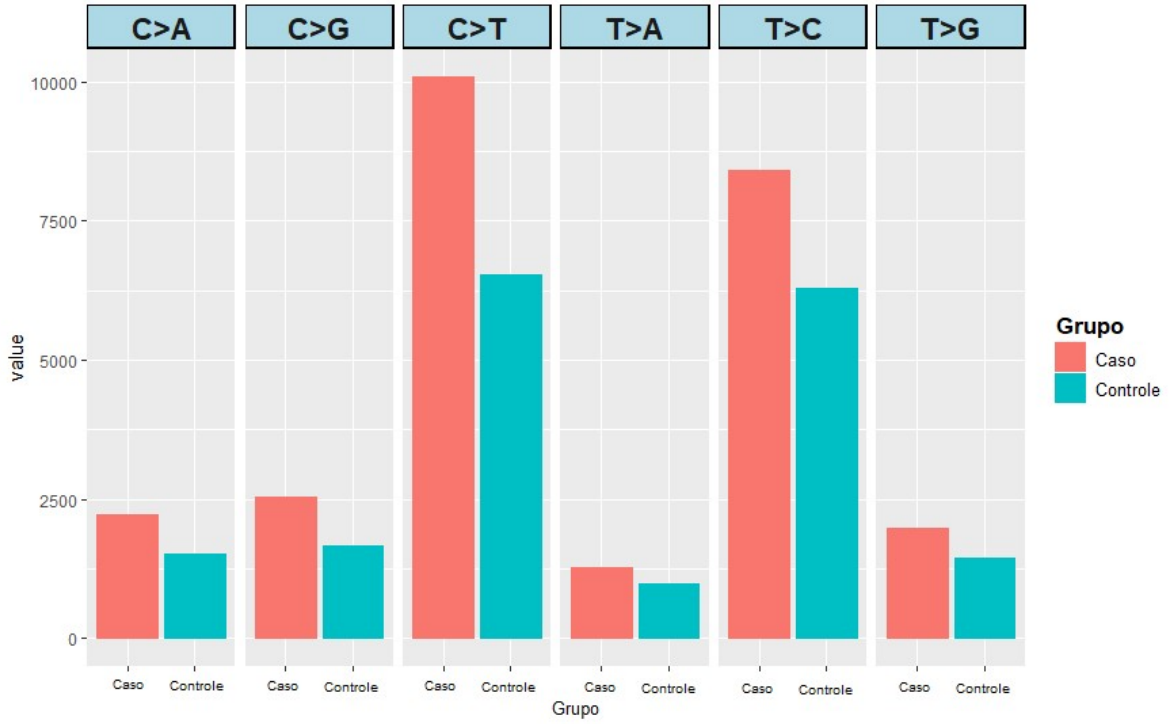
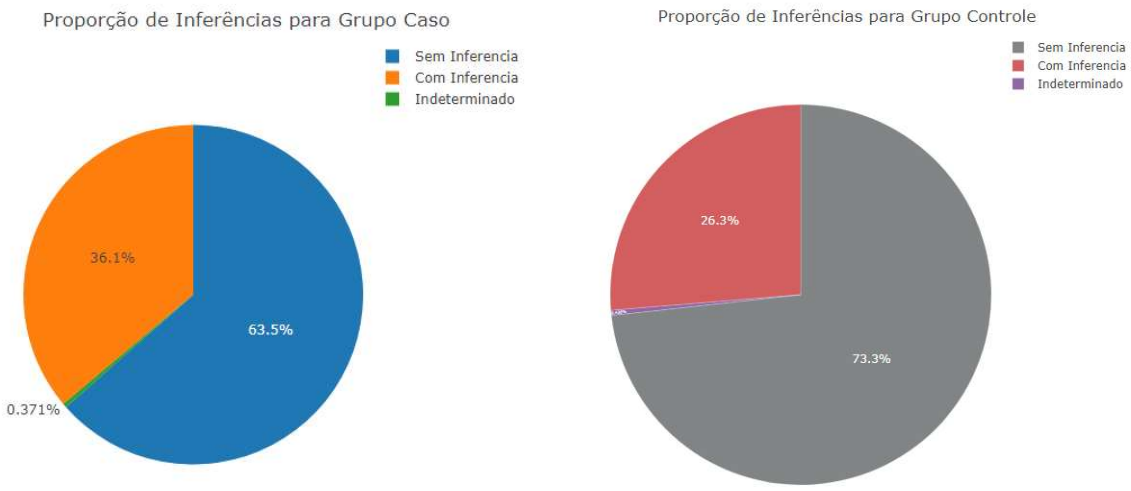


Figura 21. Mutações *de novo* em fase baseadas em alelos parentais e derivados distribuídos em classes de substituições de bases para casos e controles da população de Goiânia exposta acidentalmente a baixas doses de radiação ionizante.

No presente estudo foram identificados 36% (9.522) e 26% (5.013) de DM cujas origens não puderam ser identificadas nos grupos caso e controle, respectivamente. Nestas situações, as deduções incorporadas no SIPO permitiram inferir 4.274 e 1.995 mutações de origem materna para caso e controle, respectivamente. Por outro lado, as deduções no SIPO permitiram incluir 5.278 e 2.826 mutações de origem paterna para os grupos caso e controle, respectivamente. No entanto, 0,8% (178) de DM ainda não puderam ter sua origem estabelecida, pois as deduções no SIPO encontraram valores idênticos (Figura 22).

243

a) b)



244 **Figura 22.** Indica a proporção de inferências realizadas nos grupos caso (a) e controle (b), pois as origens dos
245 progenitores em relação às proles eram identificáveis.
246
247

6. DISCUSSÃO

Em nosso estudo, o grupo caso apresentou uma exposição à radiação com dosagens (0,2 Gy – 0,5 Gy), de acordo com o Sistema Internacional de Medidas utiliza-se a unidade Gy (Gray), que equivale a 100 rad, como unidade de dose absorvida, adotada para qualquer tipo de radiação ionizante (BIRAL, 2002). Segundo os critérios de dosimetria citogenética, as doses menores que 0,15 Gy são classificadas como muito baixas; de 0,15 a 0,3 Gy de baixas doses; de 0,31 a 0,9 de doses médias e ≥1,0 Gy são classificadas como altas doses (IAEA, 1988). Sabendo que grupos: casos e controle são da mesma população, sofrem os mesmos efeitos biossistema, conduzindo que a diferença significante entre as amostras se dá pela exposição à radiação.

A exposição parental à radiação ionizante aumenta a frequência de mutações na linha germinativa detectáveis na próxima geração (DUBROVA, 2000; LUKE, et a. 1997; BURRUEL, *et al.,* 1997; UNSCEAR, 1993), como também, que mutações *de novo* sugerem um papel importante em formas raras e comuns de doenças do neurodesenvolvimento, incluindo deficiência intelectual, autismo e esquizofrenia, tornando significativa a herdabilidade das doenças genéticas durante as gerações (VELTMAN & BRUNNER, 2012).

Estudos genômicos fornecem grandes volumes de dados que permite determinar as relações entre informações genotípicas e fenotípicas, bem como a identificação de SNPs relacionados às doenças (DENVER, *et al.,* 2009; SHAH & KUSIAK, 2004; ZANG, *et al.,* 2004). A formatação dos estudos genômicos buscam abordar a identificação de padrões, explorando, principalmente, a ciência da mineração de dados e algoritmos genéticos, outras técnicas como: árvore de decisão, heurística baseada em correlação são empregados para selecionar genes significativos. Sendo que, durante a análise desses dados pode levar a padrões de genes / SNP que podem ser responsáveis por doenças comuns, bem como pelo risco genético (SHAH & KUSIAK, 2004).

O Suíte CytoScan HD™ possui um algoritmo intrínseco, que permite a análise de um segmento cromossômico, dada a presença de marcadores polimórficos naquela região. No estudo atual, o desafio foi estabelecer o progenitor de origem para uma mutação pontual com base apenas nas transmissões de Mendel. Neste caso, a solução implantada para os achados de desvio mendeliano *de novo* foi a criação de algoritmo (SIPO), que visa a comparação de dados extraídos nos resultados obtidos pela ferramenta ChAS®, pois permitiu extrair informações relevantes do trio (pai, mãe e filho). Sendo possível exibir os

dados de marcadores para um único cromossomo. Estes resultados são extraídos e manipulados por algoritmos para realizar inferências quanto à origem da mutação e o tipo de substituição na base de DNA ocorrido.

O presente estudo realizou uma análise de desvio mendeliano para estimar a frequência de mutação *de novo* germinativa em uma população humana exposta acidentalmente às doses baixas de radiação ionizante de césio-137.

Estudo analisou os critérios de qualidade dos dados observando as recomendações exigidas pelo Fabricante (MAPD >0,25; Waviness SD >0,12; SNPQC >15), sendo que as três variáveis citadas devem ser atendidas. Realizamos um teste com 7 trios representados como grupo controle, em que o SNPQC foi inferior a 15. Estas amostras foram possíveis identificar as CNVs, porém, ao rodar o nosso *pipeline* para inferências de mutações, identificamos que a quantidade de desvios mendelianos ($\mu_{DM} \pm 0,2\%$) foi maior que a média de DM ($\mu_{DM} \sim 0,1\%$) encontradas em uma determinada população controle (Saunders et al., 2008), como também, maior que a média ($\mu_{DM} \sim 0,16\%$) encontrada no grupo exposto, demonstrando que para a condição de montagem de SNP, a métrica SNPQC é fator determinante.

Foi investigado a condição de prevalência de desvios mendelianos dentro de CNVs. Um outro algoritmo ($SIPO^{+}$) foi desenvolvido e aplicado para realizar a busca de desvios mendelianos dentro de CNVs. A conclusão é que número de mutações *de novo* foram extremamente baixo, sem significância para o estudo.

A mutação está correlacionada com o comprimento relativo dos alelos, isto é, alelos mais longos têm mais probabilidade de sofrer mutação em comparação com os mais curtos no mesmo local (GE, *et al.,* 2009). Em nossa investigação, identificamos que as mutações encontradas no grupo caso e controle foram distribuídas seguindo esta orientação, sendo que maior número de DM foram localizados em cromossomos maiores.

Até o momento, há ampla evidência que apoia as diferenças de sexo nas frequências de mutação, com os machos férteis mais velhos podendo contribuir mais para um risco de saúde mutacional do que as fêmeas mais velhas. Um maior número de divisões celulares contínuas na linha germinal masculina foi implicado como uma explicação razoável para tal diferença no efeito da idade paterna (GAO *et al*., 2019; JÓNSSON *et al*., 2017).

No entanto, nosso estudo falhou em detectar o efeito da idade materna no número de DMs. Embora tenha havido evidências crescentes de contribuições maternas para as mutações pontuais *de novo* na prole (GOLDMANN *et al*., 2016; WONG *et al*., 2016). Até o momento, há um debate contínuo sobre as contribuições maternas e paternas para a carga de mutação da linha germinativa na prole (SÉGUREL *et al*., 2014). Novas ferramentas genômicas e estatísticas aplicadas a grandes e diversos conjuntos de dados populacionais em breve ajudarão a encontrar uma solução para esse enigma biológico. Embora um número maior de trios familiares possa ser necessário para avaliar a contribuição feminina nas mutações de ponto da linha germinativa em sua prole, nossos resultados sugeriram que a força do viés de mutação masculina pode ser observada mesmo em pequenas coortes familiares.

Trabalhos publicados sobre tipos de substituições em bases nitrogenadas na cadeira de DNA estão bem representadas na linha germinativa e ocorrem espontaneamente (BROVARETS & HOVORUN, *et al.,* 2015). As substituições de base única têm sido um evento mutacional comum e frequente subjacente às divisões celulares espontaneamente que aumentam como consequência de erros de replicação do DNA ou induzidas por estressores ambientais, como RI.

Em nosso estudo foi identificado uma maior proporção entre as mutações de substituições, correspondendo por 63,92% no grupo exposto; e 73,40% no grupo controle. Em ambos os grupos a base alélica GG apresentou maior proporção de mutação *de novo* em ambos os casos; 15,76% no caso e 17,87% no controle.

As taxas de substituição de transições são mais altas (DA CRUZ, 1997) do que o esperado por acaso em relação às transversões (LYONS & LAURING, 2017). Em nossos achados, foi possível confirmar que existe uma predominância significativa na proporção de tipo de substituição por transição em relação ao tipo de substituição por transversão, tanto para grupo exposto, quanto para o grupo para controle.

Todos os SNPs que abrigam transições C> T nos conjuntos de dados não foram localizados nas ilhas CpG e foram incluídos nas análises. RI é conhecido por causar freios de suporte duplo e todos os tipos de substituições de bases, favorecendo a conhecida hipótese de que o genoma humano abriga um viés mutacional em direção à composição A / T no estande do DNA (LYNCH, 2010). Em nosso estudo, embora a linha de base do FM_{DM} em SNPs fosse diferente, os espectros mutacionais de casos e controles, considerando

todas as substituições de base, eram notavelmente semelhantes. Esta observação apoia reivindicações anteriores sobre o efeito aleatório da deposição de energia de radiação em sistemas biológicos (Vértes *et al*., 2003).

A taxa de recombinação é mais alta para mulheres do que homens, e os filhos de mães mais velhas têm mais recombinações maternas que as de mães jovens. No entanto, os homens transmitem um número muito maior de mutações aos filhos do que as mulheres (KONG, *et al*., 2012; KONG, *et al*., 2004). Nosso estudo revelou que a porcentagem de desvio mendelianos tem uma proporção maior na origem paterna do que a origem materna, porém a diferença não foi significativa.

O aumento das mutações com a idade dos pais se manifesta principalmente, talvez inteiramente, no cromossomo herdado do pai e o número de mutações *de novo* gerado a partir do espermatozoide, que aumenta a chance de uma criança sofrer uma mutação deletéria (não necessariamente limitada a mutações SNP), podendo desencadear uma maior probabilidade de autismo ou esquizofrenia (FRANCIOLI, *et al.,* 2015; FRANCIOLI, *et al,* 2014; MICHAELSON, *et al.,* 2012; KONG, *et al*., 2012; KONDRASHOV, 2003). Para nosso estudo, utilizamos somente a idade paterna como referência, pois trabalhamos com a projeção da variável condicionante o número de meioses paternas para casos e controles em função da idade dos pais à época da concepção da prole.

As deduções incorporadas no SIPO de DM cujas origens não puderam ser identificadas nos grupos caso e controle, foi numericamente pouco representativo, neste estudo, por métodos previamente estabelecidos. Nestas situações, o SIPO se apresenta como um algoritmo novo, eficaz para ser usado como ferramenta adicional na definição de origem parental de variantes polimórficas obtidas pelos microarranjos de SNPs.

Este estudo é pioneiro na avaliação dos DM na prole de um grupo de indivíduos acidentalmente expostos à RI pelo Cesio-137 usando a Plataforma Thermo Fisher. Além disso, o único estudo *in vivo* que relata o uso de pequenas CNVs para estimar a taxa de mutação germinativa de novo em humanos de uma prole de indivíduos acidentalmente expostos à radiação ionizante pelo Césio-137 foi também realizado pelo nosso grupo (COSTA *et al*., 2018).

Os DM ocorreram aleatoriamente entre os grupos casos e controle, sendo que ambos apresentaram uma baixa frequência de chamadas. Não foram identificados

marcadores com DM que se repetiam em um número significativo, mostrando a confiabilidade das análises.

Para validar os achados do presente estudo, que analisou a FM_{DM} de uma coorte discreta de crianças concebidas após seus progenitores terem sido expostos acidentalmente à RI de césio-137, os autores recomendam que estudos envolvendo coortes maiores, contendo diferentes intensidade de doses absorvidas, decorrentes de exposição terapêutica ou ocupacional, sejam desenvolvidos para avaliar a real contribuição do desvios mendelianos como marcadores retrospectivos de exposição à RI em populações humanas.

7. CONCLUSÃO

Este estudo foi pioneiro na análise de DM, correspondendo a variação de SNP polimórficos, como biomarcador de exposição útil para estimar retrospectivamente a taxa de mutação germinativa *de novo* em uma população humana exposta acidentalmente às doses baixas de radiação ionizante de césio-137 e estabelecer o *burden* das mutações germinativas na progênie.

Descobrimos que o sexo dos progenitores não teve efeito sobre o FM_{DM} de SNPs autossômicos, para ambos os grupos de caso e controle, a mãe e os pais contribuíram com números iguais de DM *de novo* para seus filhos. Depois de contabilizar a idade, nosso grupo de controle revelou que os pais mais velhos contribuíram com mais DM para seus filhos, o que poderia ser modelado pelo número de divisões de espermatogônias mitóticas em função da idade, apoiando achados anteriores de viés de mutação masculina. No entanto, nosso estudo não identificou, de forma significative, detecção do efeito da idade materna na frequência de DM.

Em síntese, houve um aumento de 44% na FM_{DM} da geração F1 de indivíduos expostos acidentalmente à RI de Césio-137, com doses absorvidas variando de 0,2 a 0,5Gy, durante o acidente radiológico de Goiânia.

Portanto, com a incorporação das deduções para tomada de decisão sobre a origem dos DM na prole, o SIPO foi capaz de eficientemente identificar por inferência a origem parental dos desvios observados, como também, indicar o tipo de substituição de bases no presente estudo.

Nesse contexto, estudos futuros envolvendo o comportamento de DM frente aos agravos genômico e mutagênico causado pela exposição e agentes ambientais poderão fornecer conhecimentos importantes sobre os efeitos biológicos, riscos e mecanismos subjacentes à exposição humana a tais agentes.

8. REFERÊNCIA BIBLIOGRÁFICA

1. ABDI Hervé, WILLIAMS Lynne J. **Main Component Analysis**. John Wiley e Sons, Inc. WIREs Comp Stat2. 2010;2:433-59p. https://doi.org/10.1002/wics.101
2. ABELSON, Harold; DISESSA, Andrea A. **Turtle geometry: The computer as a medium for exploring mathematics**. MIT press, 1986.
3. ADEWOYE, Adeolu B. et al. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. **Nature communications**, v. 6, n. 1, p. 1-8, 2015.
4. AHO, Alfred V..; JOHNSON, Stephen C.. . LR parsing. **ACM Computing Surveys (CSUR)**, v. 6, n. 2, p. 99-124, 1974.
5. ALKURAYA, Fowzan S. Discovery of mutations for Mendelian disorders. **Human genetics**, v. 135, n. 6, p. 615-623, 2016.
6. AMARAL, Andreia J. et al. Linkage disequilibrium decay and haplotype block structure in the pig. **Genetics**, v. 179, n. 1, p. 569-579, 2008.
7. AMBROS, Inge M. et al. Ultra-high density SNParray in neuroblastoma molecular diagnostics. **Frontiers in oncology**, v. 4, p. 202, 2014.
8. AMRHEIN, Valentin; GREENLAND, Sander; MCSHANE, Blake. Scientists rise up against statistical significance. 2019.
9. AMUNDSON, Sally A. et al. Biological indicators for the identification of ionizing radiation exposure in humans. **Expert Review of Molecular Diagnostics**, v. 1, n. 2, p. 211-219, 2001.
10. AZZATO, Elizabeth M. et al. A genome-wide association study of prognosis in breast cancer. **Cancer Epidemiology and Prevention Biomarkers**, v. 19, n. 4, p. 1140-1143, 2010.
11. BARBOSA, Luís Felipe FM et al. Machine learning methods applied to drilling rate of penetration prediction and optimization-A review. **Journal of Petroleum Science and Engineering**, v. 183, p. 106332, 2019.
12. BAROSS, Ágnes et al. Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. **BMC bioinformatics**, v. 8, n. 1, p. 368, 2007.
13. BARRETT, Jeffrey C. et al. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, v. 21, n. 2, p. 263-265, 2005.
14. BARTLETT, Maurice Stevenson. Properties of sufficiency and statistical tests. **Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences**, v. 160, n. 901, p. 268-282, 1937.
15. BERNHARD, Eric J. et al. Effects of ionizing radiation on cell cycle progression. **Radiation and environmental biophysics**, v. 34, n. 2, p. 79-83, 1995.
16. BETANCUR, Catalina. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. **Brain research**, v. 1380, p. 42-77, 2011.
17. BIRAL, Antônio Renato. Radiações ionizantes para médicos, físicos e leigos. In: **Radiações ionizantes para médicos, fisicos e leigos**. 2002. p. 230-230.
18. BOUQUET, A.; SØRENSEN, A. C.; JUGA, J. Genomic selection strategies to optimize the use of multiple ovulation and embryo transfer schemes in dairy cattle breeding programs. **Livestock Science**, v. 174, p. 18-25, 2015.
19. BRENNER, David J. et al. Cancer risks attributable to low doses of ionizing

radiation: assessing what we really know. **Proceedings of the National Academy of Sciences**, v. 100, n. 24, p. 13761-13766, 2003.
20. BROVARETS', Ol'ha O.; HOVORUN, Dmytro M. Proton tunneling in the A· T Watson-Crick DNA base pair: myth or reality?. **Journal of Biomolecular Structure and Dynamics**, v. 33, n. 12, p. 2716-2720, 2015.
21. BURRUEL, Victoria R.; RAABE, Otto G.; WILEY, Lynn M. In vitro fertilization rate of mouse oocytes with spermatozoa from the F1 offspring of males irradiated with 1.0 Gy 137Cs γ-rays. **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis**, v. 381, n. 1, p. 59-66, 1997.
22. CAETANO, Alexandre Rodrigues. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, n. SPE, p. 64-71, 2009.
23. CALUS, Mario PL; VANDENPLAS, Jérémie. SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. **Genetics Selection Evolution**, v. 50, n. 1, p. 34, 2018.
24. CAMBIEN, François et al. Sequence diversity in 36 candidate genes for cardiovascular disorders. **The American Journal of Human Genetics**, v. 65, n. 1, p. 183-191, 1999.
25. CARGILL, Michele et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. **Nature genetics**, v. 22, n. 3, p. 231-238, 1999.
26. CARTER, Nigel P. Methods and strategies for analyzing copy number variation using DNA microarrays. **Nature genetics**, v. 39, n. 7, p. S16-S21, 2007.
27. CARVALHO, Sérgio; WEBER, C. Raciocínio Lógico Simplificado–Vol. I. **Editora Campus**, 2010.
28. CIOPPI, Francesca; CASAMONTI, Elena; KRAUSZ, Csilla. Age-dependent de novo mutations during spermatogenesis and their consequences. In: **Genetic Damage in Human Spermatozoa**. Springer, Cham, 2019. p. 29-46.
29. CLARK, Taane G. et al. Finding associations in dense genetic maps: a genetic algorithm approach. **Human heredity**, v. 60, n. 2, p. 97-108, 2005.
30. CLARK, Taane G.; DE IORIO, Maria; GRIFFTHS, Robert C. An evolutionary algorithm to find associations in dense genetic maps. **IEEE transactions on evolutionary computation**, v. 12, n. 3, p. 297-306, 2008.
31. CLAYTON, David; LEUNG, Hin-Tak. An R package for analysis of whole-genome association studies. **Human heredity**, v. 64, n. 1, p. 45-51, 2007.
32. COHEN, Jacob. A power primer. **Psychological bulletin**, v. 112, n. 1, p. 155, 1992.
33. COHEN, Jacob. **Statistical power analysis for the behavioral sciences**. Academic press, 2013.
34. COLLINS, Francis S. Of needles and haystacks: finding human disease genes by positional cloning. **Clin Res**, v. 39, p. 615-623, 1991.
35. COLLINS, Francis S.; GUYER, Mark S.; CHAKRAVARTI, Aravinda. Variations on a theme: cataloging human DNA sequence variation. **Science**, v. 278, n. 5343, p. 1580-1581, 1997.
36. CONTI, Simon L.; EISENBERG, Michael L. Paternal aging and increased risk of congenital disease, psychiatric disorders, and cancer. **Asian journal of andrology**, v. 18, n. 3, p. 420, 2016.
37. CORMEN, Thomas H. et al. **Introduction to algorithms**. MIT press, 2009.

38. COSTA, Emília Oliveira Alves et al. Small de novo CNVs as biomarkers of parental exposure to low doses of ionizing radiation of caesium-137. **Scientific reports**, v. 8, n. 1, p. 1-13, 2018.
39. COSTA, Emília Oliveira Alves et al. The effect of low-dose exposure on germline microsatellite mutation rates in humans accidentally exposed to caesium-137 in Goiânia. **Mutagenesis**, v. 26, n. 5, p. 651-655, 2011.
40. COULTER, Michael E. et al. Chromosomal microarray testing influences medical management. **Genetics in Medicine**, v. 13, n. 9, p. 770-776, 2011.
41. CROW, James F. The origins, patterns and implications of human spontaneous mutation. **Nature Reviews Genetics**, v. 1, n. 1, p. 40-47, 2000.
42. CUMMING, Geoff; FINCH, Sue. A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. **Educational and Psychological Measurement**, v. 61, n. 4, p. 532-574, 2001.
43. DA CRUZ, A. D. et al. Human micronucleus counts are correlated with age, smoking, and cesium-137 dose in the Goiania (Brazil) radiological accident. **Mutation Research/Environmental Mutagenesis and Related Subjects**, v. 313, n. 1, p. 57-68, 1994.
44. DA CRUZ, A. D.; GLICKMAN, B. W. Nature of mutation in the human hprt gene following in vivo exposure to ionizing radiation of cesium-137. **Environmental and molecular mutagenesis**, v. 30, n. 4, p. 385-395, 1997.
45. DA CRUZ, A. D. et al. Monitoring hprt mutant frequency over time in T-lymphocytes of people accidentally exposed to high doses of ionizing radiation. **Environmental and molecular mutagenesis**, v. 27, n. 3, p. 165-175, 1996.
46. DA CRUZ, A. D. et al. Microsatellite mutations in the offspring of irradiated parents 19 years after the Cesium-137 accident. **Mutation Research/Genetic Toxicology and Environmental Mutagenesis**, v. 652, n. 2, p. 175-179, 2008.
47. DARLINGTON, Richard B. Multiple regression in psychological research and practice. **Psychological bulletin**, v. 69, n. 3, p. 161, 1968.
48. DE KLUIVER, Hilde et al. Paternal age and psychiatric disorders: A review. **American Journal of Medical Genetics Part B: Neuropsychiatric Genetics**, v. 174, n. 3, p. 202-213, 2017.
49. DE PAULA CRUZ, Athamy Sarah et al. Análise de Marcadores Microssatélites Localizados no Cromossomo Y (Y-STR) de Indivíduos Expostos ao Césio-137. **Revista EVS-Revista de Ciências Ambientais e Saúde**, v. 37, n. 6, 2010.
50. DE SMITH, Adam J. et al. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. **Human molecular genetics**, v. 16, n. 23, p. 2783-2794, 2007.
51. DE SOUZA TEODORO, Lilian et al. Análise por Microarranjo no Probando com Transtorno do Espectro Autista com CNV de perda em 15q11-13 e CNV de ganho em 6p27 e 22q11. 23. **Semina: Ciências Biológicas e da Saúde**, v. 38, n. 1supl, p. 91, 2018.
52. DENVER, Dee R. et al. A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. **Proceedings of the National Academy of Sciences**, v. 106, n. 38, p. 16310-16314, 2009.
53. DISKIN, Sharon J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. **Nucleic acids research**, v. 36, n. 19, p. e126-e126, 2008.

54. DU, Feng-Xing; CLUTTER, Archie C.; LOHUIS, Michael M. Characterizing linkage disequilibrium in pig populations. **International journal of biological sciences**, v. 3, n. 3, p. 166, 2007.
55. DUAN, Shiwei et al. FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. **Bioinformation**, v. 3, n. 3, p. 139, 2008.
56. DUBROVA, Yuri E. et al. Transgenerational mutation by radiation. **Nature**, v. 405, n. 6782, p. 37-37, 2000.
57. DUIJVESTEIJN, Naomi et al. A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. **BMC genetics**, v. 11, n. 1, p. 42, 2010.
58. ECKEL-PASSOW, Jeanette E. et al. Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. **BMC bioinformatics**, v. 12, n. 1, p. 220, 2011.
59. ELLEDGE, Stephen J. Cell cycle checkpoints: preventing an identity crisis. **Science**, v. 274, n. 5293, p. 1664-1672, 1996.
60. ERICKSON, Jeff. Algorithms. 1999.
61. FAN, Yong et al. Developmental potential of human oocytes reconstructed by transferring somatic cell nuclei into polyspermic zygote cytoplasm. **Biochemical and biophysical research communications**, v. 382, n. 1, p. 119-123, 2009.
62. FAZEL, Reza et al. Exposure to low-dose ionizing radiation from medical imaging procedures. **New England Journal of Medicine**, v. 361, n. 9, p. 849-857, 2009.
63. FEOFILOFF, Paulo. Minicurso de Análise de Algoritmos. 2011.
64. FLAKUS, Franz-Nikolaus. Radiation in perspective: Improving comprehension of risks. In: **Fuel and Energy Abstracts**. 1995. p. 465.
65. FOMBONNE, Eric. Epidemiology of pervasive developmental disorders. **Pediatric research**, v. 65, n. 6, p. 591-598, 2009.
66. FRANCIOLI, Laurent C. et al. Genome-wide patterns and properties of de novo mutations in humans. **Nature genetics**, v. 47, n. 7, p. 822, 2015.
67. FRANCIOLI, Laurent C. et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. **Nature genetics**, v. 46, n. 8, p. 818, 2014.
68. FRANKENBERG-SCHWAGER, M. Induction, repair and biological relevance of radiation-induced DNA lesions in eukaryotic cells. **Radiation and environmental biophysics**, v. 29, n. 4, p. 273-292, 1990.
69. FULLERTON, Stephanie M. et al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. **The American journal of human genetics**, v. 67, n. 4, p. 881-900, 2000.
70. GAO, Ziyue et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. **Proceedings of the National Academy of Sciences**, v. 116, n. 19, p. 9491-9500, 2019.
71. GARDNER, Martin J.; ALTMAN, Douglas G. Confidence intervals rather than P values: estimation rather than hypothesis testing. **Br Med J (Clin Res Ed)**, v. 292, n. 6522, p. 746-750, 1986.
72. GE, Jianye et al. Mutation rates at Y chromosome short tandem repeats in Texas populations. **Forensic Science International: Genetics**, v. 3, n. 3, p. 179-184, 2009.

73. GIANNOULATOU, Eleni et al. Contributions of intrinsic mutation rate and selfish selection to levels of de novo HRAS mutations in the paternal germline. **Proceedings of the National Academy of Sciences**, v. 110, n. 50, p. 20152-20157, 2013.
74. GIBBONS, Robert D.; HEDEKER, Donald R.; DAVIS, John M. Estimation of effect size from a series of experiments involving paired comparisons. **Journal of Educational Statistics**, v. 18, n. 3, p. 271-279, 1993.
75. GOMEZ, Emilio et al. Development of an image analysis system to monitor the retention of residual cytoplasm by human spermatozoa: correlation with biochemical markers of the cytoplasmic space, oxidative stress, and sperm function. **Journal of andrology**, v. 17, n. 3, p. 276-287, 1996.
76. GORLOV, Ivan P. et al. How to get the most from microarray data: advice from reverse genomics. **BMC genomics**, v. 15, n. 1, p. 223, 2014.
77. GURLEY, Lawrence R.; WALTERS, Ronald A.; TOBEY, Robert A. The metabolism of histone fractions: Phosphorylation and synthesis of histones in late G1-arrest. **Archives of biochemistry and biophysics**, v. 164, n. 2, p. 469-477, 1974.
78. HALUSHKA, Marc K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. **Nature genetics**, v. 22, n. 3, p. 239-247, 1999.
79. HEDGES, Larry V.; OLKIN, Ingram. **Statistical methods for meta-analysis**. Academic press, 2014.
80. HEWITT, Carl. Viewing control structures as patterns of passing messages. **Artificial intelligence**, v. 8, n. 3, p. 323-364, 1977.
81. HILL, W. G.; ROBERTSON, Alan. Linkage disequilibrium in finite populations. **Theoretical and applied genetics**, v. 38, n. 6, p. 226-231, 1968.
82. HOGEWEG, Paulien. The roots of bioinformatics in theoretical biology. **PLoS Comput Biol**, v. 7, n. 3, p. e1002021, 2011.
83. HOGEWEG, P.; HESPER, B. Interactive instruction on population interactions. **Computers in biology and medicine**, v. 8, n. 4, p. 319-327, 1978.
84. HOGEWEG, Pauline. Simulating the growth of cellular forms. **Simulation**, v. 31, n. 3, p. 90-96, 1978.
85. HOLLAND, John Henry et al. **Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence**. MIT press, 1992.
86. HU, Jiaqiao et al. A model reference adaptive search method for stochastic global optimization. **Communications in Information & Systems**, v. 8, n. 3, p. 245-276, 2008.
87. HUSSON, François; LÊ, Sébastien; PAGÈS, Jérôme. **Exploratory multivariate analysis by example using R**. CRC press, 2017.
88. HUYGHE, Jeroen R. et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. **Nature genetics**, v. 45, n. 2, p. 197-201, 2013.
89. IAFRATE, A. John et al. Detection of large-scale variation in the human genome. **Nature genetics**, v. 36, n. 9, p. 949-951, 2004.
90. IAFRATE, A. John et al. Detection of large-scale variation in the human genome. **Nature genetics**, v. 36, n. 9, p. 949-951, 2004.
91. ILIAKIS, George et al. DNA damage checkpoint control in cells exposed to ionizing radiation. **Oncogene**, v. 22, n. 37, p. 5834-5847, 2003.

92. INTERNATIONAL ATOMIC ENERGY AGENCY (IAEA). **The Radiological Accident in Goiânia**. IAEA, Vienna, 1988. 1-157 p.
93. JOMBART, Thibaut; AHMED, Ismaïl. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. **Bioinformatics**, v. 27, n. 21, p. 3070-3071, 2011.
94. JÓNSSON, Hákon et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. **Nature**, v. 549, n. 7673, p. 519-522, 2017.
95. KAO, Gary D. et al. Inhibition of phosphatidylinositol-3-OH kinase/Akt signaling impairs DNA repair in glioblastoma cells following ionizing radiation. **Journal of Biological Chemistry**, v. 282, n. 29, p. 21206-21212, 2007.
96. KASSAMBARA, Alboukadel. **Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra**. STHDA, 2017.
97. KENNEDY, Giulia C. et al. Large-scale genotyping of complex DNA. **Nature biotechnology**, v. 21, n. 10, p. 1233-1237, 2003.
98. KERKHOF, Hanneke JM et al. A genome-wide association study identifies an osteoarthritis susceptibility locus on chromosome 7q22. **Arthritis & Rheumatism: Official Journal of the American College of Rheumatology**, v. 62, n. 2, p. 499-510, 2010.
99. KHATKAR, Mehar S. et al. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. **BMC genomics**, v. 9, n. 1, p. 187, 2008
100. KNAUFF, Erik AH et al. Genome-wide association study in premature ovarian failure patients suggests ADAMTS19 as a possible candidate gene. **Human reproduction**, v. 24, n. 9, p. 2372-2378, 2009.
101. KOED, Karen et al. High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. **Cancer research**, v. 65, n. 1, p. 34-45, 2005.
102. KOGGE, Peter M.; STONE, Harold S. A parallel algorithm for the efficient solution of a general class of recurrence equations. **IEEE transactions on computers**, v. 100, n. 8, p. 786-793, 1973.
103. KOLKMAN, Judith M. et al. Single nucleotide polymorphisms and linkage disequilibrium in sunflower. Genetics, v. 177, n. 1, p. 457-468, 2007.
104. KONG, Augustine et al. Rate of de novo mutations and the importance of father's age to disease risk. **Nature**, v. 488, n. 7412, p. 471-475, 2012.
105. KONG, S. W. et al. Peripheral blood gene expression signature differentiates children with autism from unaffected siblings. **Neurogenetics**, v. 14, n. 2, p. 143-152, 2013.
106. KORBEL, Jan O. et al. Paired-end mapping reveals extensive structural variation in the human genome. **Science**, v. 318, n. 5849, p. 420-426, 2007.
107. KORN, Joshua M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. **Nature genetics**, v. 40, n. 10, p. 1253, 2008.
108. LANDER, Eric S. The new genomics: global views of biology. **Science**, v. 274, n. 5287, p. 536-539, 1996.
109. LANDI, Maria Teresa et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. **The american journal of human genetics**, v. 85, n. 5, p. 679-691, 2009.
110. LARMER, S. G.; SARGOLZAEI, M.; SCHENKEL, F. S. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. **Journal of dairy science**, v. 97, n. 5, p. 3128-3141,

2014.
111. LI, Heng. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. **Bioinformatics**, v. 27, n. 21, p. 2987-2993, 2011.
112. LI, Pei et al. An overview of SNP interactions in genome-wide association studies. **Briefings in functional genomics**, v. 14, n. 2, p. 143-155, 2015.
113. LI, Wen-Hsiung; SADLER, Lori A. Low nucleotide diversity in man. **Genetics**, v. 129, n. 2, p. 513-523, 1991.
114. LI, Zhiqiang et al. A genome-wide association study reveals association between common variants in an intergenic region of 4q25 and high-grade myopia in the Chinese Han population. **Human molecular genetics**, v. 20, n. 14, p. 2861-2868, 2011.
115. LIPSHUTZ, Robert J. et al. High density synthetic oligonucleotide arrays. **Nature genetics**, v. 21, n. 1, p. 20-24, 1999.
116. LÜCKE-HUHLE, C. et al. Comparative study of G2 delay and survival after 241 Americium-α and ^{60}Cobalt-γ irradiation. **Radiation and environmental biophysics**, v. 20, n. 3, p. 171-185, 1982.
117. LUKE, Garry A.; RICHES, Andrew C.; BRYANT, Peter E. Genomic instability in haematopoietic cells of F1 generation mice of irradiated male parents. **Mutagenesis**, v. 12, n. 3, p. 147-152, 1997.
118. LUO, Chenglong et al. Genome-wide association study of antibody response to Newcastle disease virus in chicken. **BMC genetics**, v. 14, n. 1, p. 42, 2013.
119. LYBÆK, Helle et al. An 8.9 Mb 19p13 duplication associated with precocious puberty and a sporadic 3.9 Mb 2q23. 3q24. 1 deletion containing NR4A2 in mentally retarded members of a family with an intrachromosomal 19p-into-19q between-arm insertion. **European journal of human genetics**, v. 17, n. 7, p. 904-910, 2009.
120. LYNCH, Michael. Rate, molecular spectrum, and consequences of human mutation. **Proceedings of the National Academy of Sciences**, v. 107, n. 3, p. 961-968, 2010.
121. LYONS, Daniel M.; LAURING, Adam S. Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. **Molecular biology and evolution**, v. 34, n. 12, p. 3205-3215, 2017.
122. LYSSENKO, Valeriya; GROOP, Leif. Genome-wide association study for type 2 diabetes: clinical applications. **Current opinion in lipidology**, v. 20, n. 2, p. 87-91, 2009.
123. MAECHLER, M., ROUSSEEUW, P., STRUYF, A., HUBERT, M., HORNIK, K. **cluster: Cluster Analysis Basics and Extensions**. R package version 2.1.0. 2019
124. MAITY, Amit; MCKENNA, W. Gillies; MUSCHEL, Ruth J. The molecular basis for cell cycle delays following ionizing radiation: a review. **Radiotherapy and oncology**, v. 31, n. 1, p. 1-13, 1994.
125. Manual do Enterprise Affymetrix® Chromosome Analysis Suite 2.0 TM Software User. **Chromosome Analysis Suite 3.2 (ChAS 3.2)**. USER GUIDE. Publication Number 702943. Revision 11. ©2017 Thermo Fisher Scientific Inc.
126. MARIONI, John C. et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. **Genome biology**, v. 8, n. 10, p. R228, 2007.
127. MCELROY, Jude J. et al. Maternal coding variants in complement receptor 1 and spontaneous idiopathic preterm birth. **Human genetics**, v. 132, n. 8, p. 935-942, 2013.

128. MEGENS, Hendrik-Jan et al. Comparison of linkage disequilibrium and haplotype diversity on macro-and microchromosomes in chicken. **BMC genetics**, v. 10, n. 1, p. 86, 2009.
129. METTLER JR, Fred A. et al. Medical radiation exposure in the US in 2006: preliminary results. **Health physics**, v. 95, n. 5, p. 502-507, 2008.
130. MICHAELSON, Jacob J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. **Cell**, v. 151, n. 7, p. 1431-1442, 2012.
131. MINSKY, Marvin; PAPERT, Seymour A. **Perceptrons: An introduction to computational geometry**. MIT press, 2017.
132. MONARD, Maria Carolina; NICOLETTI, Maria do Carmo; NOGUCHI, Raul Hideo. O cálculo proposicional: uma abordagem voltada à compreensão da linguagem Prolog. **Notas Didáticas do ICMSC-USP**, n. 5, p. 62, 1992.
133. MOSIER, Charles I. I. Problems and designs of cross-validation 1. **Educational and Psychological Measurement**, v. 11, n. 1, p. 5-11, 1951.
134. MOSLEY, Jonathan D. et al. Mechanistic phenotypes: an aggregative phenotyping strategy to identify disease mechanisms using GWAS data. **PloS one**, v. 8, n. 12, 2013.
135. MOTSINGER-REIF, Alison A. et al. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. **BMC research notes**, v. 1, n. 1, p. 65, 2008.
136. MOURA, Ronald et al. Exome analysis of HIV patients submitted to dendritic cells therapeutic vaccine reveals an association of CNOT1 gene with response to the treatment. **Journal of the International AIDS Society**, v. 17, n. 1, p. 18938, 2014.
137. MUDUNURI, Uma et al. bioDBnet: the biological database network. **Bioinformatics**, v. 25, n. 4, p. 555-556, 2009.
138. NAKAMURA, Nori et al. Radiation effects on human heredity. **Annual review of genetics**, v. 47, p. 33-50, 2013.
139. NATARAJAN, A. T. et al. A cytogenetic follow-up study of the victims of a radiation accident in Goiania (Brazil). **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis**, v. 247, n. 1, p. 103-111, 1991.
140. NAZEER, Ahsan; GHAZIUDDIN, Mohammad. Autism spectrum disorders: clinical features and diagnosis. **Pediatric Clinics of North America**, v. 59, n. 1, p. 19-25, ix, 2012.
141. NICKERSON, Deborah A. et al. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. **Nature genetics**, v. 19, n. 3, p. 233-240, 1998.
142. NIELSEN, Bent. Bartlett correction of the unit root test in autoregressive models. **Biometrika**, v. 84, n. 2, p. 500-504, 1997.
143. OKUNO, Emico. Efeitos biológicos das radiações ionizantes: acidente radiológico de Goiânia. **Estudos avançados**, v. 27, n. 77, p. 185-200, 2013.
144. PAPERT, Seymour. Children, computers and powerful ideas. 1990.
145. PEIFFER, Daniel A. et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. **Genome research**, v. 16, n. 9, p. 1136-1148, 2006.
146. PEIXOTO, Renato Mesquita et al. Evaluation of Solvent Toxicity of Plant Extract with Antiviral Action in Refrigerated Goat Semen. **Acta Scientiae Veterinariae**, v. 45, n. 1, p. 8, 2017.

147. PENROSE, L. R. Parental age and mutation. **Lancet**, v. 269, p. 312-313, 1955.

148. PETERSEN, Gloria M. et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22. 1, 1q32. 1 and 5p15. 33. **Nature genetics**, v. 42, n. 3, p. 224, 2010.
149. PFEUFER, Arne et al. Genome-wide association study of PR interval. **Nature genetics**, v. 42, n. 2, p. 153, 2010.
150. PHILLIPS, Patrick C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. **Nature Reviews Genetics**, v. 9, n. 11, p. 855-867, 2008.
151. PINKEL, Daniel et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. **Nature genetics**, v. 20, n. 2, p. 207-211, 1998.
152. PORTO-NETO, Laercio R.; KIJAS, James W.; REVERTER, Antonio. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. **Genetics Selection Evolution**, v. 46, n. 1, p. 22, 2014.
153. PURCELL, Shaun et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American journal of human genetics**, v. 81, n. 3, p. 559-575, 2007
154. R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing**, Vienna, Austria. 2018.
155. RAMALHO, A. T.; NASCIMENTO, A. C. H.; NATARAJAN, A. T. Dose assessments by cytogenetic analysis in the Goiania (Brazil) radiation accident. **Radiation Protection Dosimetry**, v. 25, n. 2, p. 97-100, 1988.
156. REDON, Richard et al. Global variation in copy number in the human genome. **nature**, v. 444, n. 7118, p. 444-454, 2006.
157. RIEDER, Mark J. et al. Sequence variation in the human angiotensin converting enzyme. **Nature genetics**, v. 22, n. 1, p. 59-62, 1999.

158. RISCH, N.; MERIKANGAS, K. **The future of genetic studies of complex human diseases**. Science, 1996. 1516–1517 p.

159. RODIG, Scott J. et al. The pre-B-cell receptor associated protein VpreB3 is a useful diagnostic marker for identifying c-MYC translocated lymphomas. **haematologica**, v. 95, n. 12, p. 2056-2062, 2010.
160. ROSENBLATT, Frank. **Principles of neurodynamics. perceptrons and the theory of brain mechanisms**. Cornell Aeronautical Lab Inc Buffalo NY, 1961.
161. ROYALL, Richard M. The effect of sample size on the meaning of significance tests. **The American Statistician**, v. 40, n. 4, p. 313-315, 1986.
162. ROYSTON, Patrick. Remark AS R94: A remark on algorithm AS 181: The W-test for normality. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 44, n. 4, p. 547-551, 1995.
163. SABETI, Pardis C. et al. Genome-wide detection and characterization of positive selection in human populations. **Nature**, v. 449, n. 7164, p. 913-918, 2007.
164. SAUNDERS, Ian W.; BROHEDE, Jesper; HANNAN, Garry N. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. **Genomics**, v. 90, n. 3, p. 291-296, 2007.
165. SEBAT, Jonathan et al. Large-scale copy number polymorphism in the human genome. **Science**, v. 305, n. 5683, p. 525-528, 2004.

166. SEDDON, Johanna M. et al. Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. **Nature genetics**, v. 45, n. 11, p. 1366, 2013.
167. SHAH, Shital C.; KUSIAK, Andrew. Data mining and genetic algorithm based gene/SNP selection. **Artificial intelligence in medicine**, v. 31, n. 3, p. 183-196, 2004.
168. SHI, Lingling et al. Whole-genome sequencing in an autism multiplex family. **Molecular autism**, v. 4, n. 1, p. 8, 2013.
169. SKANDALIS, Adonis et al. Molecular analysis of T-lymphocyte HPRT-mutations in individuals exposed to ionizing radiation in Goiânia, Brazil. **Environmental and molecular mutagenesis**, v. 29, n. 2, p. 107-116, 1997.
170. SOUTH, Sarah T. et al. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. **Genetics in Medicine**, v. 15, n. 11, p. 901-909, 2013.
171. STANKIEWICZ, Paweł; LUPSKI, James R. Structural variation in the human genome and its role in disease. **Annual review of medicine**, v. 61, p. 437-455, 2010.
172. STERNE, Jonathan AC; SMITH, George Davey. Sifting the evidence—what's wrong with significance tests?. **Physical therapy**, v. 81, n. 8, p. 1464-1469, 2001.
173. SULLIVAN, Gail M.; FEINN, Richard. Using effect size—or why the P value is not enough. **Journal of graduate medical education**, v. 4, n. 3, p. 279-282, 2012.
174. SUTHERLAND, Betsy M. et al. Clustered DNA damages induced in isolated DNA and in human cells by low doses of ionizing radiation. **Proceedings of the National Academy of Sciences**, v. 97, n. 1, p. 103-108, 2000.
175. SUZUKI, Masao et al. Radiation-quality dependent cellular response in mutation induction in normal human cells. **Journal of radiation research**, p. 0908070107-0908070107, 2009.
176. TANHA, Kiarash; MOHAMMADI, Neda; JANANI, Leila. P-value: What is and what is not. **Medical journal of the Islamic Republic of Iran**, v. 31, p. 65, 2017.
177. TATSUMOTO, Shoji et al. Direct estimation of de novo mutation rates in a chimpanzee parent-offspring trio by ultra-deep whole genome sequencing. **Scientific reports**, v. 7, n. 1, p. 1-12, 2017.
178. TOBEY, Robert A. Different drugs arrest cells at a number of distinct stages in G_2. **Nature**, v. 254, n. 5497, p. 245-247, 1975.
179. TORRES, Santiago P.; CASTRO, Carlos A. Parallel particle swarm optimization applied to the static transmission expansion planning problem. In: **2012 Sixth IEEE/PES Transmission and Distribution: Latin America Conference and Exposition (T&D-LA)**. IEEE, 2012. p. 1-6.
180. TOYOKUNI, Hideaki et al. The contribution of radiation-induced large deletion of the genome to chromosomal instability. **Radiation research**, v. 171, n. 2, p. 198-203, 2009.
181. TUCKER, Tracy et al. Comparison of genome-wide array genomic hybridization platforms for the detection of copy number variants in idiopathic mental retardation. **BMC medical genomics**, v. 4, n. 1, p. 25, 2011.
182. TUZUN, Eray et al. Fine-scale structural variation of the human genome. **Nature genetics**, v. 37, n. 7, p. 727-732, 2005.
183. UNSCEAR. **Hereditary Effects of Radiation United Nations.** United Nations, New York, 2001.
184. UNSCEAR. **Sources and Effects of Ionizing Radiation.** United Nations, New York, 1993.

185. VARGA, Elizabeth A. et al. The prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. **Genetics in Medicine**, v. 11, n. 2, p. 111-117, 2009.
186. VASSON, Aurélie et al. Custom oligonucleotide array-based CGH: a reliable diagnostic tool for detection of exonic copy-number changes in multiple targeted genes. **European Journal of Human Genetics**, v. 21, n. 9, p. 977-987, 2013.
187. VELTMAN, Joris A.; BRUNNER, Han G. De novo mutations in human genetic disease. **Nature Reviews Genetics**, v. 13, n. 8, p. 565-575, 2012.
188. VERMEESCH, Joris Robert et al. Guidelines for molecular karyotyping in constitutional genetic diagnosis. **European Journal of Human Genetics**, v. 15, n. 11, p. 1105-1114, 2007.
189. VÉRTES, Attila et al. (Ed.). **Handbook of Nuclear Chemistry: Vol. 1: Basics of Nuclear Science; Vol. 2: Elements and Isotopes: Formation, Transformation, Distribution; Vol. 3: Chemical Applications of Nuclear Reactions and Radiation; Vol. 4: Radiochemistry and Radiopharmaceutical Chemistry in Life Sciences; Vol. 5: Instrumentation, Separation Techniques, Environmental Issues; Vol. 6: Nuclear Energy Production and Safety Issues**. Springer Science & Business Media, 2010.
190. VILLAR, Bruno. **Raciocínio lógico completo.** Ed. Método. 2016. 4; São Paulo. ISBN: 978-85-309-6838-0
191. WAIN, Louise V.; ARMOUR, John AL; TOBIN, Martin D. Genomic copy number variation, human health, and disease. **The Lancet**, v. 374, n. 9686, p. 340-350, 2009.
192. WIGGANS, G. R. et al. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. **Journal of dairy science**, v. 92, n. 7, p. 3431-3436, 2009.
193. WINCHESTER, Laura; YAU, Christopher; RAGOUSSIS, Jiannis. Comparing CNV detection methods for SNP arrays. **Briefings in functional genomics and proteomics**, v. 8, n. 5, p. 353-366, 2009.
194. WU, Thomas D.; NACU, Serban. Fast and SNP-tolerant detection of complex variants and splicing in short reads. **Bioinformatics**, v. 26, n. 7, p. 873-881, 2010.
195. XU, Yaji et al. Genome-wide algorithm for detecting CNV associations with diseases. **BMC bioinformatics**, v. 12, n. 1, p. 331, 2011.
196. XUE, Yali et al. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. **Current Biology**, v. 19, n. 17, p. 1453-1457, 2009.
197. YANDELL, Brian S. **Practical data analysis for designed experiments**. Crc Press, 1997.
198. ZAHIR, Farah R.; MARRA, Marco A. Use of Affymetrix arrays in the diagnosis of gene copy-number variation. **Current protocols in human genetics**, v. 85, n. 1, p. 8.13. 1-8.13. 13, 2015.
199. ZHOU, Bin-Bing S.; ELLEDGE, Stephen J. The DNA damage response: putting checkpoints in perspective. **Nature**, v. 408, n. 6811, p. 433-439, 2000.

423

424

APÊNDICE I

Informações sobre as bibliotecas configuradas no ChAS®

No quadro 1 temos informações sobre a biblioteca de referência aplicada aos estudos.

Quadro 1. Informações sobre a biblioteca de referência aplicada ao estudo

Arquivo de parâmetro de análise	CytoScanHD_Array.na33.annot.db
Nome da matriz	CytoScan HD Array
Versão da ChAS	3.2
Versão do Genoma UCSC	hg19
Versão do Genoma NCBI	37
Versão do dbSNP	132

No quadro 2 consta informações sobre os dados de entrada.

Quadro 2. Informações sobre dados de entrada

Lista os arquivos contemplados na geração do Trio	F01-1F_(CytoScanHD_Array).cyhd.cychp; F01-2M_(CytoScanHD_Array).cyhd.cychp; F01-3P_(CytoScanHD_Array).cyhd.cychp
Entrada de Cromossomo	1 a 22 e os cromossos sexuais: X e Y

No quadro 3 identifica as variáveis que foram manipuladas.

Quadro 3 Informações sobre as variáveis obtidas a partir do ChAS®

Marcador	Identificador do marcador de SNP
Genotipagem	Chamada de genotipagem. Informações bialélicas. Formando três combinações: AA/BB/AB
Valor de Confiança	Valor de confiança para genotipagem
Sinal A	Valor do sinal bruto para o sinal A no marcador
Sinal B	Valor do sinal bruto para o sinal B no marcador
Base Nitrogenada	Chamada da base nitrogenada. Informações bialélicas. Formando dez combinações: AA|AC|AG|AT|CC|CG|CT|GG|GT|TT
dbSNP	Valor do identificador do SNP

Cromossomo	Cromossomo associado ao marcador
Posição do Cromossomo	Posição do Cromossomo no SNP

437
438

APÊNDICE II

Protoloco de *pipeline* para o PLINK

1. Utilizamos a base de dados (MySQL©) para extrair os dados dos SNPs dos trios do grupo caso e controle.
2. Aplicamos os filtros:
    a) Eliminamos os SNPs com *call rate* > $5x10^{-2}$;
    b) Eliminamos os SNPs em que as informações de IDSNP; Posição; Cromossomos estavam sem valores ou nulos;
    c) Eliminamos os SNPs em que as informações de genótipos do Pai e/ou Mãe e/ou Prole constavam sem valores ou nulos;
    d) Eliminamos os SNPs que geraram mutações em suas proles;
3. Após a aplicação dos filtros, foi gerado um *dataset* de 522.172 SNPs selecionados em autossomos.
4. A partir deste *dataset* foram gerados dois arquivos que são interpretados pelo PLINK.
    a) O primeiro é o arquivo .MAP, que possui a seguinte estrutura física:

Cromossomo	IDSNP	Distância Genética	Posição
1	rs2340582	0	882803
1	rs3748597	0	888659

    b) O segundo é o arquivo .PED, que possui a seguinte estrutura física:

Família	ID Indivíduo	Presença Pai*	Presença Mãe**	Sexo***	Fenótipo****	rs2340582	rs3748597
C1P	1	1	0	1	1	G G	C C
F4M	2	0	1	2	2	A G	C T

*Presença do Pai: 1 – Presença, 0 – Ausência; **Presença do Mãe: 1 – Presença, 0 – Ausência; ***Sexo: 1 – Masculino, 2 – Feminino; ****Fenótipo: 1 – Controle, 2 – Caso.

5. A opção *--file* usa um único parâmetro, a raiz dos nomes dos arquivos de entrada e procura dois arquivos: um arquivo PED e um arquivo MAP:
plink --file cesio

6. Criando um arquivo PED binário. Este formato compacta os dados economizando espaço e acelera as análises subseqüentes. Para criar um arquivo PED binário, use o seguinte comando:
plink --file cesio --make-bed --out cesio
Foram gerados os arquivos: cesio.bed; cesio.bim; cesio.fam

7. Executando o *pipeline* com o arquivo PED binário: Para especificar que os dados de entrada estão no formato binário, em oposição ao formato PED / MAP de texto normal, basta usar a opção --bfile em vez de –file:
plink –bfile cesio

8. Ligação de desequilibrio baseado no desbaste de SNP: Optamos por gerar um subconjunto com desbaste de SNPs que estão em equilíbrio de ligação correlacionados entre si. Usamos o comando:
plink --bfile cesio --indep-pairwise 500 5 0.1, --indep-pairwise se baseia apenas na correlação genotípica aos pares.

9. Após o desbaste de nosso *dataset* original com 522K SNPs, foi gerado um novo *dataset* (cesio.pruned.in) contendo 2.789 SNPs com maior frequência de correlacionamento entre os SNPs selecionados.

10. Para exclusão com SNPs vazio, aplicamos o comando:
plink --bfile cesio --mind 0.1

11. Para gerar estatísticas sobre valores vazios, utilizamos o comando:
plink --bfile cesio --missing --out miss_stat

12. Geração de estatísticas sobre frequencia dos alelos:
plink --bfile cesio --freq --nonfounders --out freq_stat

Cromossomo	idSNP	A1	A2	MAF*
7	rs7790666	G	A	0.3269
8	rs6991681	A	G	0.3077

MAF (*Minor Alleles Frequency*). É a frequência com que o segundo alelo mais comum ocorre em uma determinada população

13. Associação das variâncias entre caso e controle:
plink --bfile cesio --assoc --nonfounders --out as1

Chr	**idSNP**	**BP**	**A1**	**F_A**	**F_U**	**A2**	**CHISQ**	**P**	**OR**
7	rs7790666	159118443	G	0,25	0,38	A	2,05	0,15	0,53
8	rs6991681	146292734	A	0,36	0,26	G	1,12	0,28	0,51

Segue o *layout* dos campos:
Chr: Cromossomo; idSNP: Identificador SNP; A1: Código do alelo 1 (o alelo menor e raro, com base em toda a frequência da amostra); F_A: A frequência dessa variante nos casos; F_U: A frequência dessa variante nos controles; A2: Código para o outro alelo; CHISQ: A estatística do qui-quadrado; P: O valor da significância; OR: Taxa de probabilidade para o teste.

Os resultados mostram a variante simulada do SNP rs7790666 é realmente o SNP mais significativo da lista, com uma diferença nas frequências alélicas de 0,25 nos casos versus 0,38 nos controles.

14. Obtivemos uma lista classificada dos resultados da associação, que também inclui uma faixa de valores de significância ajustados para vários testes:
plink --bfile cesio --assoc --adjust --nonfounders --out as2

15. Análise de estratificação compreende analisar uma forma de agrupamento que emparelhe indivíduos com base na identidade genética. O comando executado é:
plink --bfile cesio --cluster --mc 2 --ppc 0.05 --out str1

Neste caso aplicamos *clustering* IBS, do inglês, *identify by states*, com o comando (--cluster), usados para dois indivíduos (--mc 2) e que qualquer par de indivíduos que tenha um valor de significância menor que 0,05 para testar se deve ou não os dois indivíduos pertencem à mesma população com base nos dados disponíveis do SNP. Segue a representação dos 4 primeiros agrupamentos

SOL-0	C1P_1 C52P_15
SOL-1	C1M_2 C70M_26
SOL-2	F7M_36 F8M_38
SOL-3	F4P_31 F21P_47

16. Análise de associação, contabilizando os clusters gerados a partir do método IBS. Para essa análise correspondente, usaremos a estatística de associação *Cochran-Mantel-Haenszel* (CMH), que testa a associação entre os SNPs condicionadas ao agrupamento gerados pelo *cluster* IBS, na etapa anterior; também incluímos a opção --*adjust* para obter uma lista ordenada dos resultados da associação CMH:
plink --bfile cesio --mh --within str1.cluster2 --adjust --out aac1

17. Solicitamos que cada cluster contenha pelo menos 1 caso e 1 controle, ou seja, para que seja informativo para associação, com a opção --cc e especifique um limite de 0,01, usando o comando --ppc:
plink --bfile cesio --cluster --cc --ppc 0.01 --out version2

18. Geramos uma visualização da subestrutura deste *dataset*, criando uma matriz de distâncias IBS pares e, em seguida, usando um pacote estatístico como R para gerar um gráfico de escala multidimensional, segue os comandos:
plink --bfile cesio --cluster --matrix --out ibd_view, foi gerado o arquivo .mibs
plink --bfile cesio --cluster --distance-matrix --nonfounders --out ibd_view, foi gerado o arquivo .mdist
Passamos para a linguaguem R para gera os gráficos multidimensionais:

a) b)

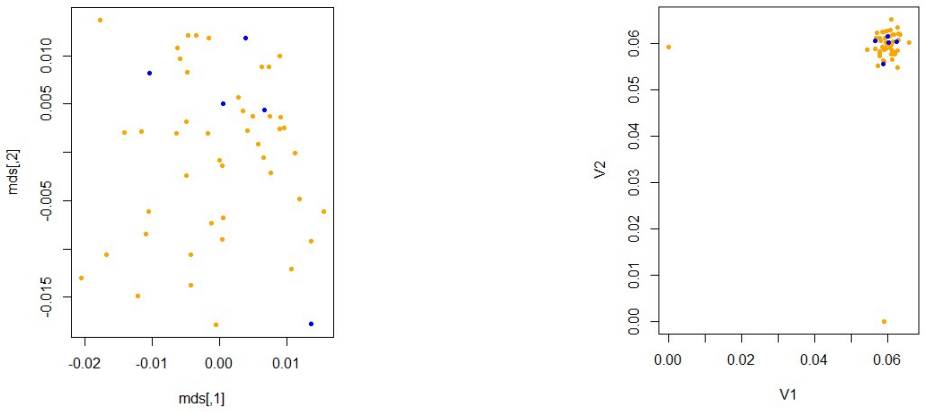


Figura 10. a) Gráfico gerado a partir do método IBS. b) Gráfico gerado a partir do método IBD. Ambos os gráficos certamente parecem sugerir que não há divergências entre os dois grupos na amostra.

19. Geramos o PCA (Análise de Componente Principal) usando o PLINK, versão 2. Executamos o comando:

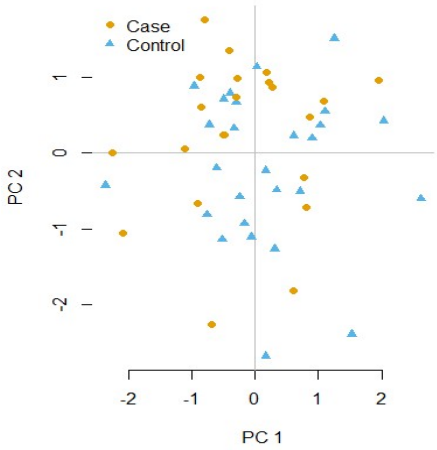
a) Geramos os arquivos binários para a montagem dos auto-valores e auto-vetores: *plink --bfile cesio --extract plink.prune.in --make-bed --out prunedcesio*

b) Utilizamos a versão 2 do PLINK, com o comando:

plink2 --bfile pruned.merge.cesio --nonfounders --pca --out pca.pruned, assim, são gerados os resultados: prunedcesio.eigenvalue e prunedcesio.eigenvector.

c) O PCA foi gerado com base no resultado prunedcesio.eigenvector, seguindo os scripts do R:

```
library(pca3d)
pca.plink<-read.table("pca.pruned.eigenvec",header = TRUE,sep = "\t",dec = ".")
pca.plink.num<-pca.plink[,-1]
stand.pca.plink<-scale(pca.plink.num)
res.pca.plink<-prcomp(stand.pca.plink, center = TRUE)
group<- factor(pca.plink[,1])
p.pca2d.plink<-pca2d(res.pca.plink, group = group, legend="topleft")
```



575 **Figura 11.** PCA de 2,7 K snps, resultado do desbaste usando a janela de 500 snps; com incrementos de 5
576 SNPs e DL máximo de 0.1. As variáveis contidas no PCA representa a matriz de relacionamento padronizada
577 por variância; dimensionamento multidimensional (MDS) com base nas distâncias de Hamming. Com base
578 na disposição dos dados neste gráfico de PCA, concluímos que não houve divergência entre os grupos na
579 amostra.

APÊNDICE III

Um documento em Markdown com *pipeline* para a geração dos resultados

````
# Um documento em Markdown para a montagem de resultados da tese final

```r
##Fluxo do algoritmo: SIPO
library(DiagrammeR)
mermaid("
   graph TB
     id1(SIPO)==>id2[MINERAÇÃO <br/> DE DADOS]
     id2==>id3(VALOR DE CONFIANÇA > 0.05)
     id3==>id4(VALIDAR <br/> LAYOUT);
     id4==>id5(ANALISAR VARIÁVEIS)
     id1==>id6[GENÓTIPOS]
     id6==>id7(GENÓTIPOS MATERNA-BN)
     id7==>id8(GENÓTIPOS PATERNA-BN)
     id8==>id9(GENÓTIPOS PROLE-BN)
     id1==>id10[DESVIO MENDELIANO]
     id10==>id11(SEM <br/>INFERÊNCIA)
     id10==>id12(COM <br/>INFERÊNCIA)
     id12==>id13(MENOR VALOR DE CONFIANÇA)
     id13==>id14(MUTAÇÃO MAIS PRÓXIMA); id14==>id15
     id11==>id15>RESULTADO: ORIGEM DA MUTAÇÃO (PROGENITORES) <br/>E TIPO DE
SUBSTITUIÇÃO]
")
```



```r
#Normalidade e Variância
  ## 1. Testes de normalidade
  dfCasoDM<-read.csv("dfCasoDM.csv",header = TRUE,sep = ",",dec = ".")
  dfControlDM<-read.csv("dfControlDM.csv",header = TRUE,sep = ",",dec = ".")
  Parâmetros:
  xbCaso <- mean(dfCasoDM$dm_caso) # média
  xbControl <- mean(dfControlDM$dm_control) # média
  meCaso <- median(dfCasoDM$dm_caso) # mediana
  meControl <- median(dfControlDM$dm_control) # mediana
  sxCaso <- round(sd(dfCasoDM$dm_caso), digits = 2) # desvio padrão
  sxControl <- round(sd(dfControlDM$dm_control), digits = 2) # desvio padrão
  cat("\n Média amostral do Grupo Caso =", xbCaso, "\n Desvio padrão amostral do Grupo Caso =", sxCaso,
"\n Mediana amostral do Grupo Caso =", meCaso)
  cat("\n Média amostral do Grupo Controle =", xbControl, "\n Desvio padrão amostral do Grupo Controle
=", sxControl, "\n Mediana amostral do Grupo Controle =", meControl)
t1Caso <- shapiro.test(rnorm(dfCasoDM$dm_caso, xbCaso, sxCaso)) # Shapiro-Wilk Caso
t1Control <- shapiro.test(rnorm(dfControlDM$dm_control, xbControl, sxControl)) # Shapiro-Wilk Controle

```

```r
##estatísticas do grupo caso
library(ggplot2)
dfIdadeGeral<-read.csv("dfIdadeGeral.csv",header = TRUE,sep = ";",dec = ".")
````
````

```
meanCaso <- round(mean(dfCasoDM$dm_caso), digits = 2)
sdCaso <- round(sd(dfCasoDM$dm_caso), digits = 2)
medianaCaso <- round(median(dfCasoDM$dm_caso), digits = 2)
somaCaso <- round(sum(dfCasoDM$dm_caso), digits = 2)
meanAgePatIdadeCase<-round(mean(dfIdadeGeral$age_pat_case), digits = 2)
meanAgeMatIdadeCase<-round(mean(dfIdadeGeral$age_mat_case), digits = 2)
sdAgePatIdadeCase<-round(sd(dfIdadeGeral$age_pat_case), digits = 2)
sdAgeMatIdadeCase<-round(sd(dfIdadeGeral$age_mat_case), digits = 2)
table(meanCaso, sdCaso, medianaCaso, somaCaso, meanAgePatIdadeCase, meanAgeMatIdadeCase,
sdAgePatIdadeCase, sdAgeMatIdadeCase)
print(table(meanCaso, sdCaso, medianaCaso, somaCaso, meanAgePatIdadeCase, meanAgeMatIdadeCase,
sdAgePatIdadeCase, sdAgeMatIdadeCase), zero.print = ".")
##estatísticas do grupo controle
dfControlDM<-read.csv("dfControlDM.csv",header = TRUE,sep = ",",dec = ".")
meanControl <- round(mean(dfControlDM$dm_control), digits = 2)
sdControl <- round(sd(dfControlDM$dm_control), digits = 2)
medianaControl <- round(median(dfControlDM$dm_control), digits = 2)
somaControl <- round(sum(dfControlDM$dm_control), digits = 2)
meanAgePatIdadeControl<-round(mean(dfIdadeGeral$age_pat_control), digits = 2)
meanAgeMatIdadeControl<-round(mean(dfIdadeGeral$age_mat_control), digits = 2)
sdAgePatIdadeControl<-round(sd(dfIdadeGeral$age_pat_control), digits = 2)
sdAgeMatIdadeControl<-round(sd(dfIdadeGeral$age_mat_control), digits = 2)
table(meanControl, sdControl, medianaControl, somaControl, meanAgePatIdadeControl,
meanAgeMatIdadeControl, sdAgePatIdadeControl, sdAgeMatIdadeControl)
print(table(meanControl, sdControl, medianaControl, somaControl, meanAgePatIdadeControl,
meanAgeMatIdadeControl, sdAgePatIdadeControl, sdAgeMatIdadeControl), zero.print = ".")
```
```r
##Hipóteses do teste de Normalidade:
##H0: Distribuição é Normal
##H1: Distribuição não é Normal
##Gráfico QQ com envelope
nControl<-length(dfControlDM$dm_control)
xbControl <- mean(dfControlDM$dm_control) # média
sxControl <- round(sd(dfControlDM$dm_control), digits = 2) # desvio padrão
cat("\n n =", nControl)
nSimControl <- 100 # Número de simulações
confControl <- 0.95 # Coef. de confiança
# Dados simulados ~ normal
dadosSimControl <- matrix(rnorm(nControl * nSimControl, mean = xbControl, sd = sxControl), nrow =
nControl)
dadosSimControl <- apply(dadosSimControl, 2, sort)
# Limites da banda e média
infSupControl <- apply(dadosSimControl, 1, quantile, probs = c((1 - confControl) / 2,
 (1 + confControl) / 2))
xbSimControl <- rowMeans(dadosSimControl)
# Gráfico
#faixayControl <- range(dfControlDM$dm_control, dadosSimControl) -18.54337 3524.30691
faixayControl <- c(-18.54337, 3524.30691) # mesmos valores do grupo caso
qqControl <- qqnorm(dfControlDM$dm_control, main = "Gráfico Q-Q: Desvio Mendeliano - Grupo
Controle", xlab = "Quantile", pch = 20,
 ylab = "Nº Desvio Mendeliano - Grupo Controle", ylim = faixayControl) # usar os mesmos valores nas
coordenadas dos eixos do grupo caso
eixoxControl <- sort(qqControl$x)
lines(eixoxControl, xbSimControl)
lines(eixoxControl, infSupControl[1,])
lines(eixoxControl, infSupControl[2,])
```

```
dadosSimControl <- matrix(rnorm(nControl * nSimControl, mean = xbControl, sd = sxControl), nrow =
nControl)
dadosSimControl <- apply(dadosSimControl, 2, sort)

```
```r
##Hipóteses do teste de Normalidade:
##H0: Distribuição é Normal
##H1: Distribuição não é Normal
p >> 0,05 → podemos assumir que os dados sigam a distribuição Normal
## 2. Gráfico QQ com envelope
nCaso<-length(dfCasoDM$dm_caso)
xbCaso <- mean(dfCasoDM$dm_caso) # média
sxCaso <- round(sd(dfCasoDM$dm_caso), digits = 2) # desvio padrão
cat("\n n =", nCaso)
nSimCaso <- 100 # Número de simulações
confCaso <- 0.95 # Coef. de confiança
# Dados simulados ~ normal
dadosSimCaso <- matrix(rnorm(nCaso * nSimCaso, mean = xbCaso, sd = sxCaso), nrow = nCaso)
dadosSimCaso <- apply(dadosSimCaso, 2, sort)
# Limites da banda e média
infSupCaso <- apply(dadosSimCaso, 1, quantile, probs = c((1 - confCaso) / 2,
 (1 + confCaso) / 2))
xbSimCaso <- rowMeans(dadosSimCaso)
# Gráfico
faixayCaso <- range(dfCasoDM$dm_caso, dadosSimCaso)
qqCaso <- qqnorm(dfCasoDM$dm_caso, main = "Gráfico Q-Q: Desvio Mendeliano - Grupo Caso", xlab =
"Quantile", pch = 20,
 ylab = "Nº Desvio Mendeliano - Grupo Caso", ylim = faixayCaso)
eixoxCaso <- sort(qqCaso$x)
lines(eixoxCaso, xbSimCaso)
lines(eixoxCaso, infSupCaso[1,])
lines(eixoxCaso, infSupCaso[2,])
dadosSimCaso <- matrix(rnorm(nCaso * nSimCaso, mean = xbCaso, sd = sxCaso), nrow = nCaso)
dadosSimCaso <- apply(dadosSimCaso, 2, sort)
```
```r
## Aplicação do Teste T (Student t)
dfdadosgerais<-read.csv("dados_gerais_caso_controle.csv",header = TRUE,sep = ";",dec = ".")
dfDadosGeraisControl<-dfdadosgerais[dfdadosgerais$Group=="Control",]
dfDadosGeraisCase<-dfdadosgerais[dfdadosgerais$Group=="Case",]
testtDM<-t.test(dfDadosGeraisCase$Total_DM, dfDadosGeraisControl$Total_DM)
testtFreqDM<-t.test(dfDadosGeraisCase$Freq_DM, dfDadosGeraisControl$Freq_DM)


## Aplicação do Teste F (Teste de variância)
testfDM<-var.test(dfDadosGeraisCase$Total_DM, dfDadosGeraisControl$Total_DM)
testfFreqDM<-var.test(dfDadosGeraisCase$Freq_DM, dfDadosGeraisControl$Freq_DM)

```

```
## Pipeline no PLINK:
## plink --file cesio --mind 0.1 --recode --out cesio
## plink --file cesio --maf 0.1 --nonfounders --noweb
## plink --file cesio --make-bed --out cesio
## plink --bfile cesio
## PRUNING
```

```
## plink --bfile cesio --indep-pairwise 500 5 0.1
## plink --bfile cesio --extract plink.prune.in --make-bed --out prunedcesio
## plink --bfile cesio --missing --out miss_stat
## Associação das variâncias entre caso e controle
## plink --bfile cesio --assoc --nonfounders --out as1
## plink --bfile cesio --assoc --adjust --nonfounders --out as2
## plink --bfile cesio --model --nonfounders --out mod1
## plink --bfile cesio --model --cell 5 --snp rs3845291 --out mod2
## plink --bfile cesio --cluster --mc 2 --ppc 0.05 --out str1
## plink --bfile cesio --mh --within str1.cluster2 --adjust --out aac1
## plink --bfile cesio --cluster --cc --ppc 0.01 --out version2
## plink --bfile cesio --mh --within version2.cluster2 --adjust --out aac2
## plink --bfile cesio --cluster --K 2 --out version3
## plink --bfile cesio --mh --adjust --out aac3
## plink --bfile cesio --cluster --matrix --out ibd_view
## plink --bfile cesio --cluster --distance-matrix --nonfounders --out ibd_view
```

```r
setwd("Drive:/../plink")
m <- as.matrix(read.table("ibd_view.mibs"))
md <- as.matrix(read.table("ibd_view.mdist"))
mds <- cmdscale(as.dist(1-m))
scale <- cmdscale(as.dist(1-md))
k <- c( rep("orange",45) , rep("blue",44) )
plot(mds,pch=20,col=k)
plot(scale,pch=20,col=k)
```

```
## Contruindo o PCA a partir dos dados gerados no pruning
## plink --bfile prunedcesio --bmerge prunedcesio.bed prunedcesio.bim prunedcesio.fam --make-bed --out
pruned.merge.cesio
## plink2 --bfile pruned.merge.cesio --nonfounders --pca --out pca.pruned
```

```r
##Gerando os gráficos de PCA a partir das informações eigenvector gerado no plink2
library(ggplot2)
library(tidyr)
library(devtools)
library(pca3d)
setwd("Drive:/../plink2")
pca.plink<-read.table("pca.pruned.eigenvec",header = TRUE,sep = "\t",dec = ".")
pca.plink.num<-pca.plink[,-1]
pca.plink.num.x<-pca.plink.num[,-1]
stand.pca.plink<-scale(pca.plink.num.x)
res.pca.plink<-prcomp(stand.pca.plink, center = TRUE)
group<- factor(pca.plink[,1])
p.pca3d.plink<-pca3d(res.pca.plink, group = group, legend="topleft")
p.pca2d.plink<-pca2d(res.pca.plink, group = group, legend="topleft")
```

```r
library(magrittr)
library(plotly)
#Plot Box-Plot para numero de DM: Caso X Controle
f1 <- list(
  family = "Arial, sans-serif",
  size = 18,
```

```
  color = "lightgrey"
)
f2 <- list(
  family = "Old Standard TT, serif",
  size = 14,
  color = "black"
)
a <- list(
  title = "Nº Desvio Mendeliano",
  titlefont = f1,
  showticklabels = TRUE,
  tickangle = 0,
  tickfont = f2,
  exponentformat = "E"
)
b <- list(
  title = "Grupo",
  titlefont = f1,
  showticklabels = TRUE,
  tickangle = 0,
  tickfont = f2,
  exponentformat = "E"
)
bpDM<-plot_ly(dfCasoDM, y = ~dfCasoDM$dm_caso, type = "box",name="Caso", boxpoints = "all", jitter
= 0.1, pointpos = -1.9)  %>%
            add_trace(dfControlDM, y = ~dfControlDM$dm_control, type = "box",name="Controle", boxpoints
= "all", jitter = 0.1, pointpos = -1.9) %>%
   layout(xaxis = b, yaxis = a, title = "Nº Desvio Mendeliano")
```
```r
library("gplots")
library(RColorBrewer)
library(tidyverse)
dfDMGrupo<-read.csv("dfDMGrupo.csv",header = TRUE,sep = ",",dec = ".")
# Cleveland dot plot
norder <- dfDMGrupo$family[order(dfDMGrupo$group, dfDMGrupo$dm)]
dfDMGrupo$family <- factor(dfDMGrupo$family, levels = norder)
pCleveland<-ggplot(dfDMGrupo, aes(x=dm, y=family)) +
            geom_segment(aes(yend=family), xend=0, colour="grey50") +
            geom_point(size=3, aes(colour=group)) +
            scale_colour_brewer(palette="Set1", limits=c("Case","Control"), guide=FALSE) +
            theme_bw() +
                theme(panel.grid.major.y = element_blank(), # No horizontal grid lines +
                    legend.position=c(1, 0.55), # Put legend inside plot area
                    legend.justification=c(1, 0.5)) +
            facet_grid(group ~ ., scales="free_y", space="free_y") +
            labs(title="Desvio Mendeliano vs Trios ~ Grupos", x="Desvio Mendeliano", y = "Trios")

```


```r
#plot das regressões linear
library("ggpubr")
library(ggplot2)
plmdosefrqdm<-ggplot(dfdadosgerais, aes(Dose, Freq_DM, colour=Dose, fill=Dose)) +
    geom_smooth(method="lm") +
    geom_point(size=3) +
    theme_bw() +
```

```
      xlab("Dose Absorvida (Gy)") +
      ylab("Frequência de Desvio Mendeliano")

```
```r
#aplicação da regressão linear da fração de todas as fases dos DMs paterna (DM Paterna/DM Materna) em relação à idade paterna
dfdadosgerais<-read.csv("dados_gerais_caso_controle.csv",header = TRUE,sep = ";",dec = ".")

p.ratio.dm.age.paternal<-ggplot(dfdadosgerais, aes(x=idade_pai, y=Razao_DM_F_M, shape=Group, colour=Group, fill=Group)) +
   geom_smooth(method="lm", fill = "grey97", size = 2, alpha = 1, se = FALSE) +
   scale_colour_manual(values = c("blue","orange")) +
   geom_point(size=3) +
   theme_bw() +
   xlab("Idade Paterna") +
   ylab("Razão de Desvios Mendelianos Paterna sobre a Materna")

```
```r
#Plot DMs paterna (Caso/Controle) em relação à idade paterna
dfDMAgeParental<-read.csv("dados_dm_age_parental.csv",header = TRUE,sep = ";",dec = ".")

#case
p.DM.Age.Parental.case<-ggplot(dfDMAgeParental, aes(x=Age_Case, y=MD_Case, shape=Parental_Case, colour= Parental_Case, fill=Parental_Case)) +
                    geom_point(size=3) +
                    facet_grid(. ~ Parental_Case) +
                    theme(strip.text = element_text(face="bold", size=rel(1.5)),
                    strip.background = element_rect(fill="lightblue", colour="black",
                    size=1)) +
                    scale_y_continuous(limits=c(0,1500),breaks = seq(0, 1500, by = 500)) +
                    labs(
                       x = "Idade Paterna - Grupo: Caso",
                       y = "Número de Desvio Mendeliano Quanto à Origem - Grupo: Caso"
                    )
#control
p.DM.Age.Parental.control<-ggplot(dfDMAgeParental, aes(x=Age_Control, y=MD_Control, shape=Parental_Control, colour= Parental_Control, fill=Parental_Control)) +
                    geom_point(size = 3) +
                    facet_grid(. ~ Parental_Control) +
                    theme(strip.text = element_text(face="bold", size=rel(1.5)),
                    strip.background = element_rect(fill="lightblue", colour="black",
                    size=1)) +
                    scale_y_continuous(limits=c(0,1500),breaks = seq(0, 1500, by = 500)) +
                    labs(
                       x = "Idade Paterna - Grupo: Controle",
                       y = "Número de Desvio Mendeliano Quanto à Origem - Grupo: Controle"
                    )

```

```r
variant_age_md<-read.csv("age_md_parental.csv",header = TRUE,sep = ";",dec = ".")
#Idade Parental vs Número de DM do grupo Controle
  p.var.age.md.control<-ggplot(variant_age_md, aes(Age_Control, MD_Control, shape=Parental_Control, colour=Parental_Control, fill=Parental_Control)) +
  geom_smooth(method="lm") +
```

```
  stat_regline_equation(
    aes(label =  paste(..eq.label.., ..rr.label.., sep = "~~~~"))
  )+
  geom_point(size=3) +
  theme_bw() +
  xlab("Idade Parental: Grupo - Controle") +
  ylab("Número de Mutações: Grupo - Controle") +
  expand_limits(y=0) +
  scale_x_continuous(limits=c(10, 60),breaks = seq(10, 60, by = 10)) +
  scale_y_continuous(limits=c(0, 1500),breaks = seq(0, 1500, by = 250))

#Idade Parental vs Número de DM do grupo Caso
  p.var.age.md.case<-ggplot(variant_age_md, aes(Age_Case, MD_Case, shape=Parental_Case,
colour=Parental_Case, fill=Parental_Case)) +
  geom_smooth(method="lm") +
  stat_regline_equation(
    aes(label =  paste(..eq.label.., ..rr.label.., sep = "~~~~"))
  )+
  geom_point(size=3) +
  theme_bw() +
  xlab("Idade Parental: Grupo - Caso") +
  ylab("Número de Mutações: Grupo - Caso") +
  expand_limits(y=0) +
  scale_x_continuous(limits=c(10, 60),breaks = seq(10, 60, by = 10)) +
  scale_y_continuous(limits=c(0, 1500),breaks = seq(0, 1500, by = 250))

```
```r
#aplicação do método anova para comparar a variável idade entre os grupos
library(ggpubr)
dfdadosgerais<-read.csv("dados_gerais_caso_controle.csv",header = TRUE,sep = ";",dec = ".")
# Compute the analysis of variance
res.aov.agepat <- aov(idade_pai ~ Group, data = dfdadosgerais)
res.aov.agemat <- aov(idade_mae ~ Group, data = dfdadosgerais)

# Summary of the analysis
summary(res.aov.agepat)
p.bp.age.pat<-ggboxplot(dfdadosgerais, x = "Group", y = "idade_pai",
          color = "Group", palette = c("#00AFBB", "#FFA500"),
          add = "jitter", shape = "Group",
          ylab = "Idade Paterna", xlab = "Grupo")

p.ggline.age.pat<-ggline(dfdadosgerais, x = "Group", y = "idade_pai",
       add = c("mean_se", "jitter","violin"),
       order = c("Case", "Control"),
       ylab = "Idade Paterna", xlab = "Grupo")

#Tukey multiple pairwise-comparisons
#### The function STARTS here ####
#Fonte: https://rpubs.com/aaronsc32/post-hoc-analysis-tukey
#age.lm<-(idade_pai ~ Group, data=dfdadosgerais)
#age.aov<-aov(age.aov)
plotTukeyHSD <- plotTukeysHSD <- function(tukey.out,
                          x.axis.label = "Comparison",
                          y.axis.label = "Effect Size",
                        axis.adjust = 0,
                        adjust.x.spacing = 5){

```

```
983     tukey.out <- as.data.frame(tukey.out[[1]])
984     means <- tukey.out$diff
985     categories <- row.names(tukey.out)
986     groups <- length(categories)
987     ci.low <- tukey.out$lwr
988     ci.up  <- tukey.out$upr
989
990     n.means <- length(means)
991
992     #determine where to plot points along x-axis
993     x.values <- 1:n.means
994     x.values <- x.values/adjust.x.spacing
995
996
997     # calculate values for plotting limits
998     y.max <- max(ci.up) +
999       max(ci.up)*axis.adjust
1000    y.min <- min(ci.low) -
1001      max(ci.low)*axis.adjust
1002
1003    if(groups == 2){ x.values <- c(0.25, 0.5)}
1004    if(groups == 3){ x.values <- c(0.25, 0.5,0.75)}
1005
1006    x.axis.min <- min(x.values)-0.05
1007    x.axis.max <- max(x.values)+0.05
1008
1009    x.limits <- c(x.axis.min,x.axis.max)
1010
1011    #Plot means
1012    plot(means ~ x.values,
1013        xlim = x.limits,
1014        ylim = c(y.min,y.max),
1015        xaxt = "n",
1016        xlab = "",
1017        ylab = "",
1018        cex = 1.25,
1019        pch = 16)
1020
1021    axis(side = 1,
1022        at = x.values,
1023        labels = categories,
1024       )
1025
1026    #Plot upper error bar
1027    lwd. <- 2
1028    arrows(y0 = means,
1029           x0 = x.values,
1030           y1 = ci.up,
1031           x1 = x.values,
1032           length = 0,
1033           lwd = lwd.)
1034
1035    #Plot lower error bar
1036    arrows(y0 = means,
1037           x0 = x.values,
1038           y1 = ci.low,
1039           x1 = x.values,
1040           length = 0,
```

```
        lwd = lwd.)

  #add reference line at 0
  abline(h = 0, col = 2, lwd = 2, lty =2)

  mtext(text = x.axis.label,side = 1,line = 1.75)
  mtext(text = y.axis.label,side = 2,line = 1.95)
  mtext(text = "Error bars = 95% CI",side = 3,line = 0,adj = 0)
}

tukey.age.pat.group<-TukeyHSD(res.aov.agepat)
par(mfrow = c(1,2))
plot(tukey.age.pat.group)
plotTukeysHSD(tukey.age.pat.group)

tukey.age.mat.group<-TukeyHSD(res.aov.agemat)
par(mfrow = c(1,2))
plot(tukey.age.mat.group)
plotTukeysHSD(tukey.age.mat.group)

library(multcomp)
glht.anova.age.pat<-summary(glht(res.aov.agepat, linfct = mcp(group = "Tukey")))

#Multiple Comparisons of Means: Tukey Contrasts
aov.age.pat<-aov(formula = idade_pai ~ Group, data = dfdadosgerais)

#Pairewise t-test
pairwise.aov.age.pat<-pairwise.t.test(dfdadosgerais$idade_pai, dfdadosgerais$Group,
            p.adjust.method = "BH")
```

```r
library(cluster)
library(fpc)
#cluster
cls.age.alelo.pai<-read.table("dataset_pca_age_alelos_pai.txt",header = TRUE,sep = "\t",dec = ".")
df.cls.age.alelo.pai <- cls.age.alelo.pai[, -1] # without known classification
df.cls.age.alelo.pai.x <- df.cls.age.alelo.pai[, -1]
# Kmeans clustre analysis
clust.age.alelo.pai <- kmeans(df.cls.age.alelo.pai.x, centers=2)
# More complex
clusplot.age.alelo.pai<- clusplot(df.cls.age.alelo.pai, clust.age.alelo.pai$cluster,
            main = 'Cluster solution: Number of Mutation Deviation Found in the Father for Alleles vs Age',
            color=TRUE,
            shade=TRUE,
            labels=5,
            lines=0)

cls.age.alelo.mae<-read.table("dataset_pca_age_alelos_mae.txt",header = TRUE,sep = "\t",dec = ".")
df.cls.age.alelo.mae.x <- cls.age.alelo.mae[, -1] # without known classification
# Kmeans clustre analysis
clust.age.alelo.mae <- kmeans(df.cls.age.alelo.mae.x, centers=2)
# More complex
clusplot.age.alelo.mae<- clusplot(df.cls.age.alelo.mae.x, clust.age.alelo.mae$cluster,
            main = 'Cluster solution: Number of Mutation Deviation Found in the Mother for Alleles vs Age',
            color=TRUE,
            shade=TRUE,
            labels=5,
```

```
            lines=0)
```
```r
library(plotly)
infer.alelo.caso<-read.csv("num_inferencia_origem_caso_2.csv",header = TRUE,sep = ";",dec = ".")
p.pie.infer.caso <- plot_ly(infer.alelo.caso, labels = ~Inference, values = ~Event, type = 'pie')
p.pie.infer.caso <- p.pie.infer.caso %>% layout(title = 'Proportion of Inferences Made for Group Case',
       xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
       yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
infer.alelo.control<-read.csv("num_inferencia_origem_controle_2.csv",header = TRUE,sep = ";",dec = ".")
colors <- c('rgb(211,94,96)', 'rgb(128,133,133)', 'rgb(144,103,167)')
p.pie.infer.control <- plot_ly(infer.alelo.control, labels = ~Inference, values = ~Event, type = 'pie',
                        textposition = 'inside',
                        textinfo = 'label+percent',
                        insidetextfont = list(color = '#FFFFFF'),
                        hoverinfo = 'text',
                        marker = list(colors = colors,
                        line = list(color = '#FFFFFF', width = 1)),
                        #The 'pull' attribute can also be used to create space between the sectors
                        showlegend = FALSE)
p.pie.infer.control <- p.pie.infer.control %>% layout(title = 'Proportion of Inferences Made for Group
Control',
       xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
       yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

```r
#aplicação do método anova para comparar a variável idade entre os grupos
library(ggpubr)
anova.nummeiosepat<-read.table("dataset_pca_num_meiose_pat_geral_2.txt",header = TRUE,sep = "\t",dec
= ".")
# Compute the analysis of variance
res.aov.nummeiosepat <- aov(num.meiose.paterna ~ Grupo, data = anova.nummeiosepat)
# Summary of the analysis
summary(res.aov.nummeiosepat)
p.bp.num.meiose.pat<-ggboxplot(anova.nummeiosepat, x = "Grupo", y = "num.meiose.paterna",
        color = "Grupo", palette = c("#00AFBB", "#FFA500"),
        add = "jitter", shape = "Grupo",
        ylab = "Número de Meiose Paterna", xlab = "Grupo")

p.ggline.num.meiose<-ggline(anova.nummeiosepat, x = "Grupo", y = "num.meiose.paterna",
      add = c("mean_se", "jitter","violin"),
      order = c("Caso", "Controle"),
      ylim = c(4.00E+010, 6.00E+11),
      ylab = "Número de Meiose Paterna", xlab = "Grupo")
```

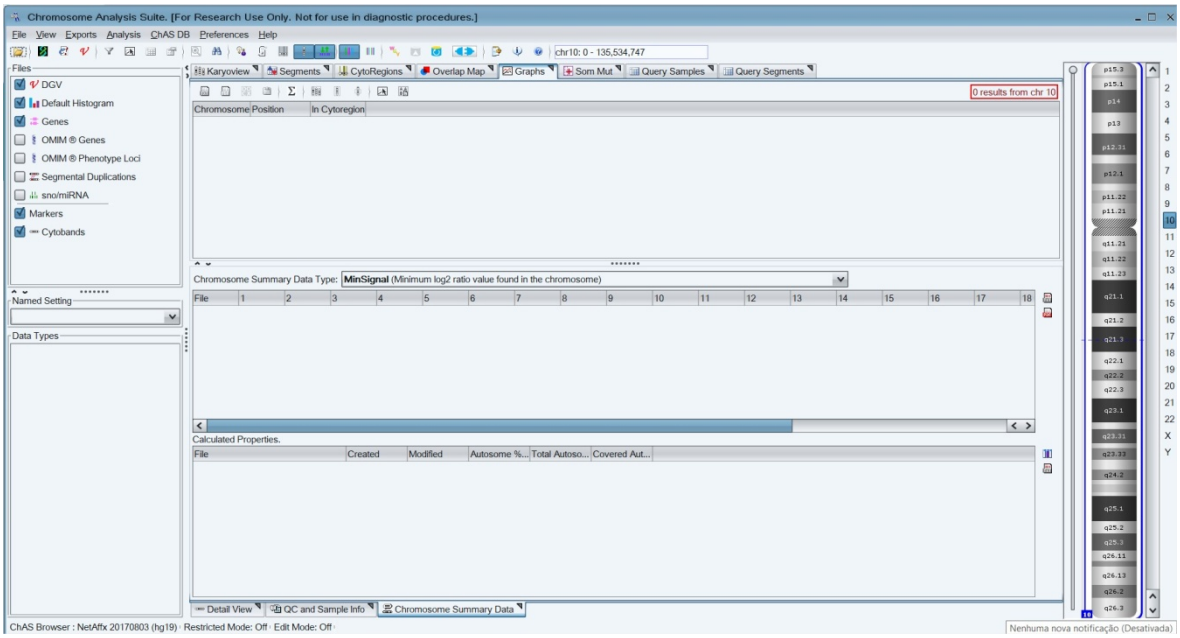
```r
library(plotly)
infer.alelo.caso<-read.csv("num_inferencia_origem_caso_tese.csv",header = TRUE,sep = ";",dec = ".")
infer.alelo.caso <- infer.alelo.caso %>%
     as.character() %>%
     stri_trans_general("Latin-ASCII") %>%
     toupper()
p.pie.infer.caso <- plot_ly(infer.alelo.caso, labels = ~Inference, values = ~Event, type = 'pie')
p.pie.infer.caso <- p.pie.infer.caso %>% layout(title = 'Proporção de Inferências para Grupo Caso',
        xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
        yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

```
infer.alelo.control<-read.csv("num_inferencia_origem_controle_tese.csv",header = TRUE,sep = ";",dec = ".")
colors <- c('rgb(211,94,96)', 'rgb(128,133,133)', 'rgb(144,103,167)')
p.pie.infer.control <- plot_ly(infer.alelo.control, labels = ~Inference, values = ~Event, type = 'pie',
                               textposition = 'inside',
                               textinfo = 'percent',
                               insidetextfont = list(color = '#FFFFFF'),
                               hoverinfo = 'text',
                               marker = list(colors = colors,
                               line = list(color = '#FFFFFF', width = 0.2)),
                               #The 'pull' attribute can also be used to create space between the sectors
                               showlegend = TRUE)
p.pie.infer.control <- p.pie.infer.control %>% layout(title = 'Proporção de Inferências para Grupo Controle',
      xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
      yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
```

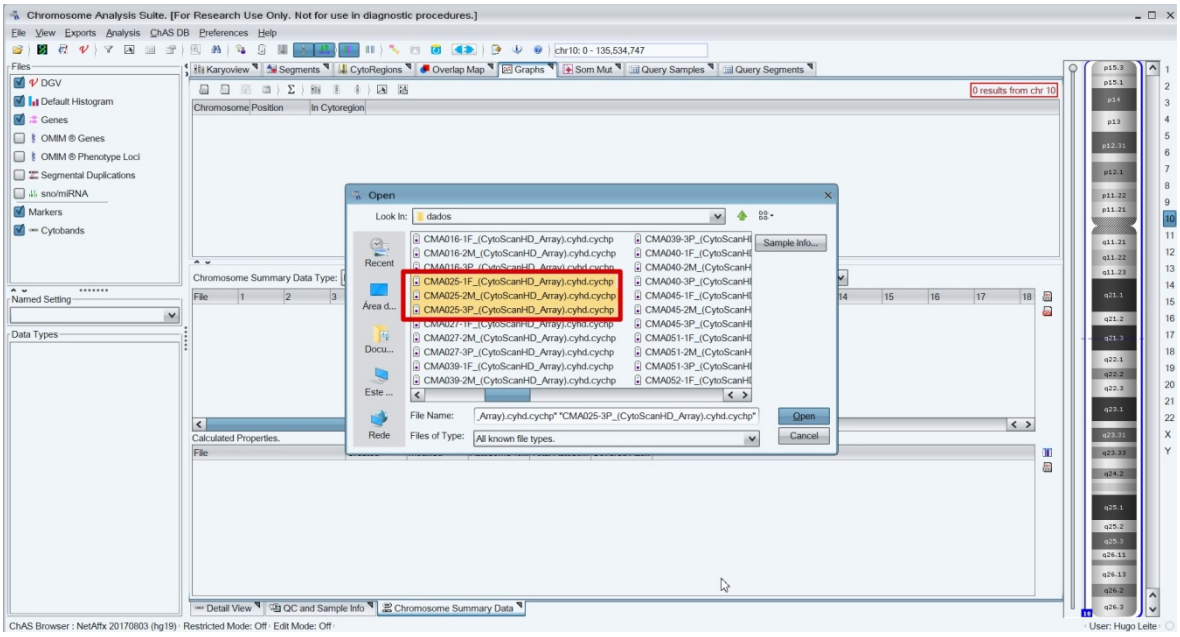
ANEXO I

Protocolo da ferramenta ChAS®

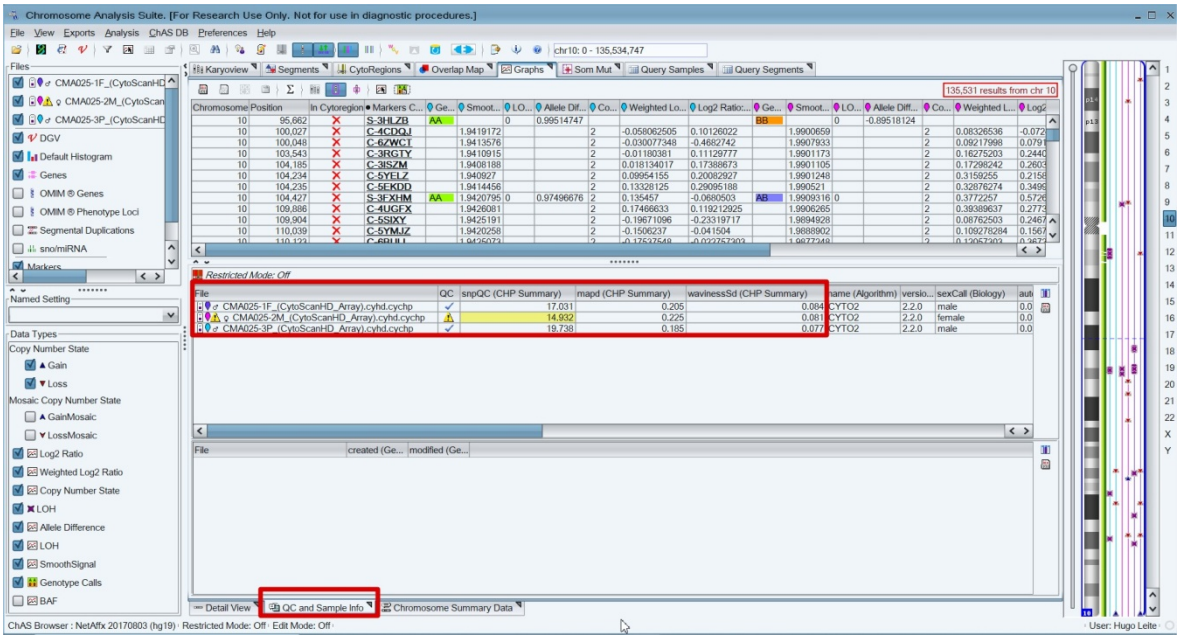
1. Tela incial da ferramenta ChAS®



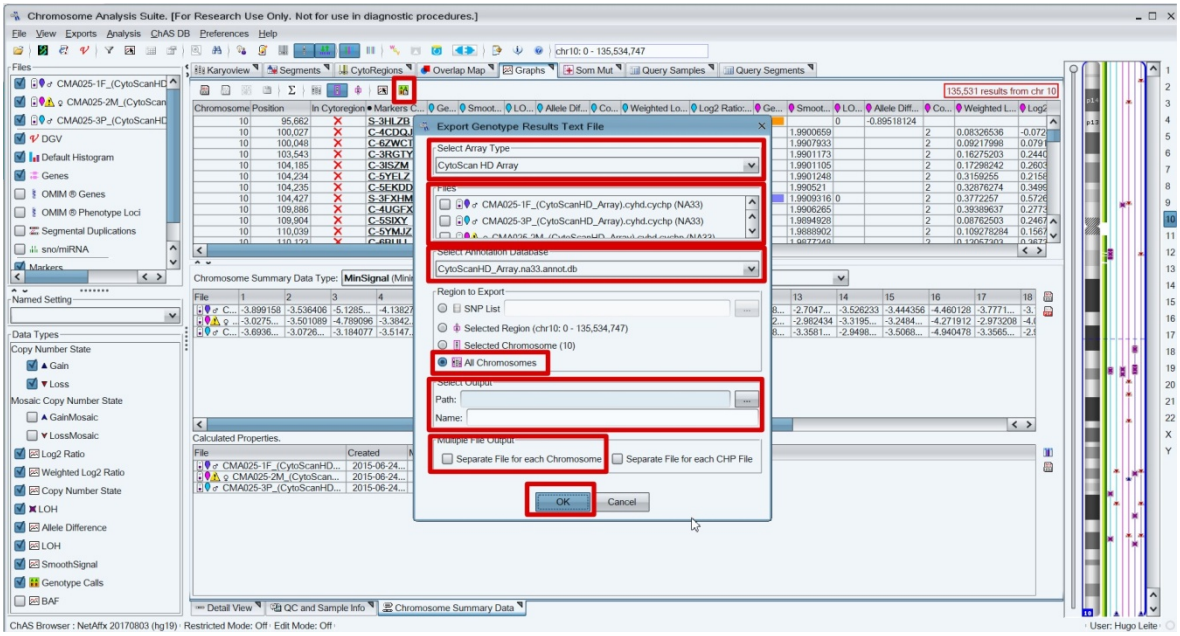
2. Carregando os dados do Trio



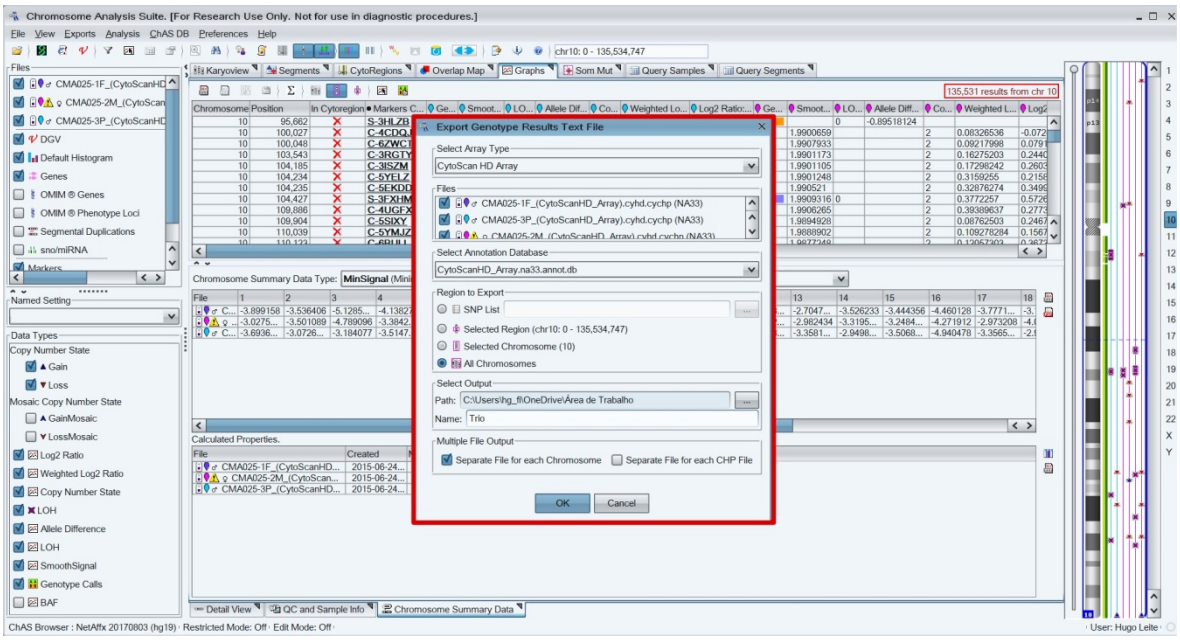
3. Verificando informações sobre o QC (Controle de Qualidade)



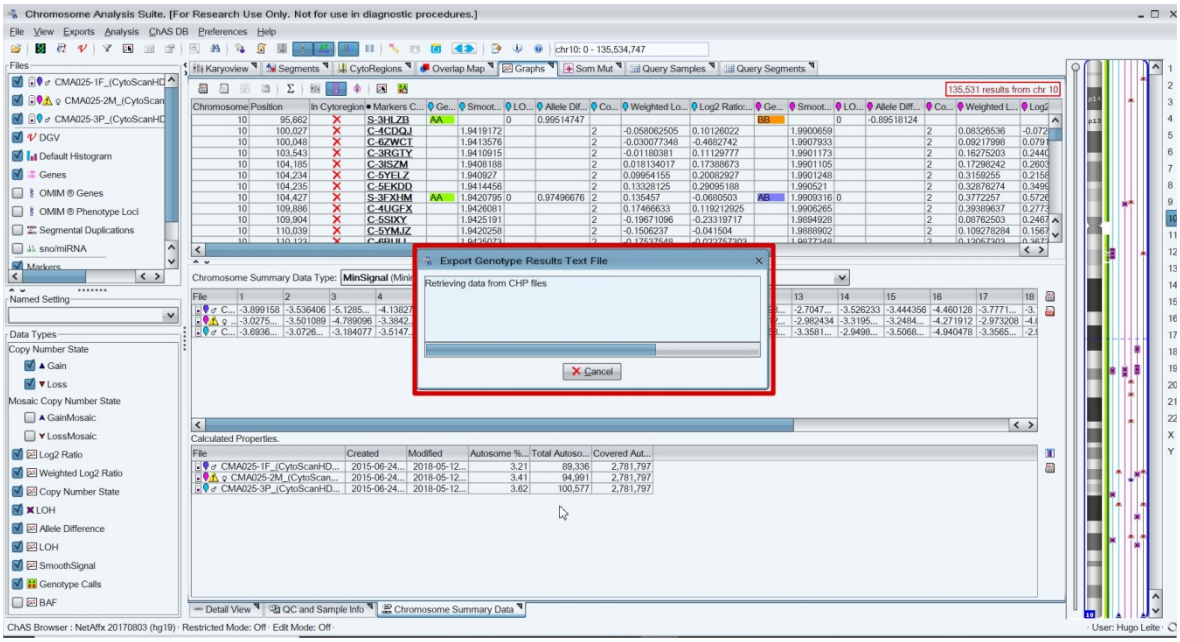
1. Exportando os resultados dos genótipos para um arquivo .txt



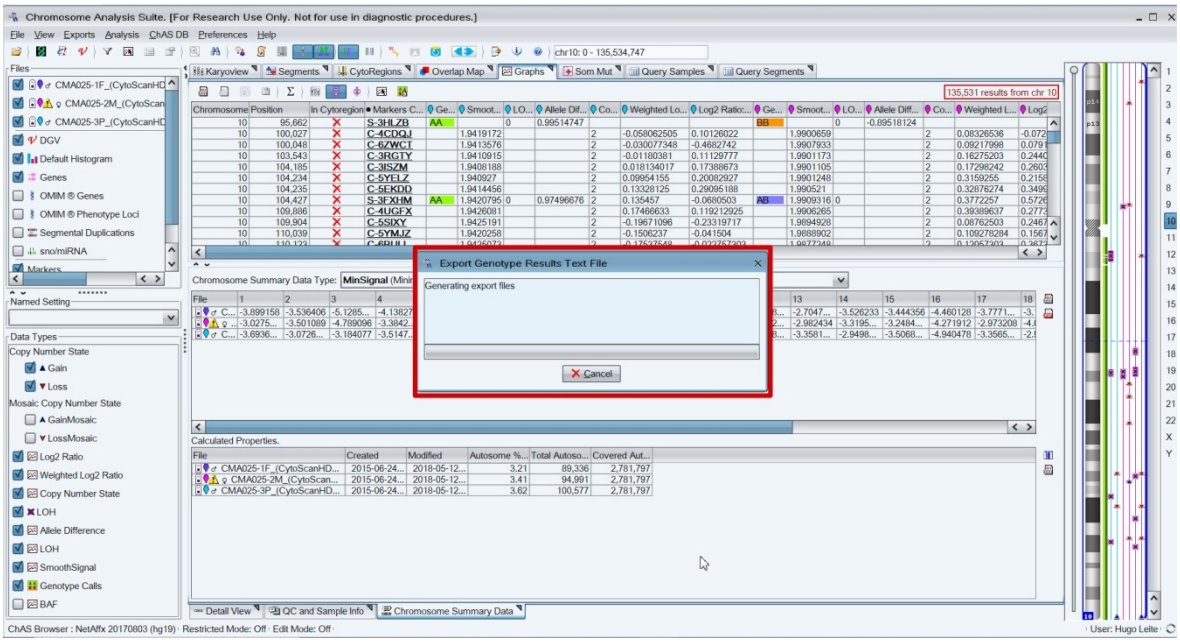
2. Informando os parâmetros para gerar os resultados dos genótipos



3. Exportando os resultados dos genótipos para o arquivo .txt



4. Gerando o arquivo .txt com os dados de genótipos exportados



ANEXO II

Biblioteca do GeneChip® CytoScan HD™ 750K

Dados da biblioteca aplicado ao GeneChip® CytoScan HD™ 750K:
Annotation DB Used: #C:\Affymetrix\ChAS\Library\CytoScanHD_Array.na32.3.annot.db # Array Type Name: CytoScan HD Array # Array Type Internal Name: CytoScanHD_Array # Export GUID: 0d2ffa6e-5aea-4acc-a462-57943082c6c2 # Array Annotation Database NetAffx Build: 32.3 # UCSC Genomic Version: hg19 # NCBI Genomic Version: 37 # dbSNP Version: 132 # CHP File 1: E:\Affymetrix\Resultados\aCGH001-1F_(CytoScanHD_Array).cyhd.cychp (NA32.3) # CHP File 2: E:\Affymetrix\Resultados\aCGH001-2M_(CytoScanHD_Array)(2).cyhd.cychp (NA32.3) # CHP File 3: E:\Affymetrix\Resultados\aCGH0013P_(CytoScanHD_Array).cyhd.cychp (NA32.3) # Input Chromosome: All # Output Chromosome: All

ANEXO III

Layout de saída do GeneChip® CytoScan HD™ 750K – Gerado pelo software ChAS

ID	Call-Codes-F	Confidence-F	Base-Calls-F	Call-Codes-M	Confidence-M	Base-Calls-M	Call-Codes-P	Confidence-P	Base-Calls-P	dbSNP	Chr	Position
S-3WRNV	BB	8.870682E-13	GG	BB	0.0	GG	BB	0.0	GG	rs2340582	1	882803
S-4GXBG	AA	8.437695E-15	CC	AA	0.0	CC	AA	2.220446E-16	CC	rs3748597	1	888659
S-3VYOT	AB	1.489919E-13	CT	AB	1.756372E-13	CT	AB	0.0	CT	rs6696609	1	903426
S-4SUCW	AA	1.997115E-6	CC	AA	3.416449E-10	CC	AA	5.639933E-13	CC	rs28695703	1	904355

Layout de saída do SIPO

Exemplos de SNPs que apresentaram desvios mendellianos

ID	Call-Codes-F	Confidence-F	Base-Calls-F	Call-Codes-M	Confidence-M	Base-Calls-M	Call-Codes-P	Confidence-P	Base-Calls-P	dbSNP	Chr	Position	Origem Mutação	TP*
S-3PZYJ	BB	3.2511274E-5	AA	AB	0.0	AG	AA	3.2511274E-5	GG	rs16839451	1	4737693	a	e
S-4NUMZ	AB	8.437695E-15	AG	BB	0.0	GG	BB	2.882139E-12	GG	rs3748597	1	888659	b	f
S-3VBSG	AB	1.489919E-13	AC	AA	0.007451657	AA	AA	0.0029024158	AA	rs6696609	1	903426	c	g
S-3NJMQ	AB	1.997115E-6	CT	AA	3.416449E-10	CC	AA	8.633136E-7	CC	rs28695703	1	904355	d	h

*TP: Tipo de Substituição

a) Origem da Mutacao: Origem Paterna;

b) Origem da Mutacao: Origem Paterna;

c) Origem da Mutacao: Origem Paterna;

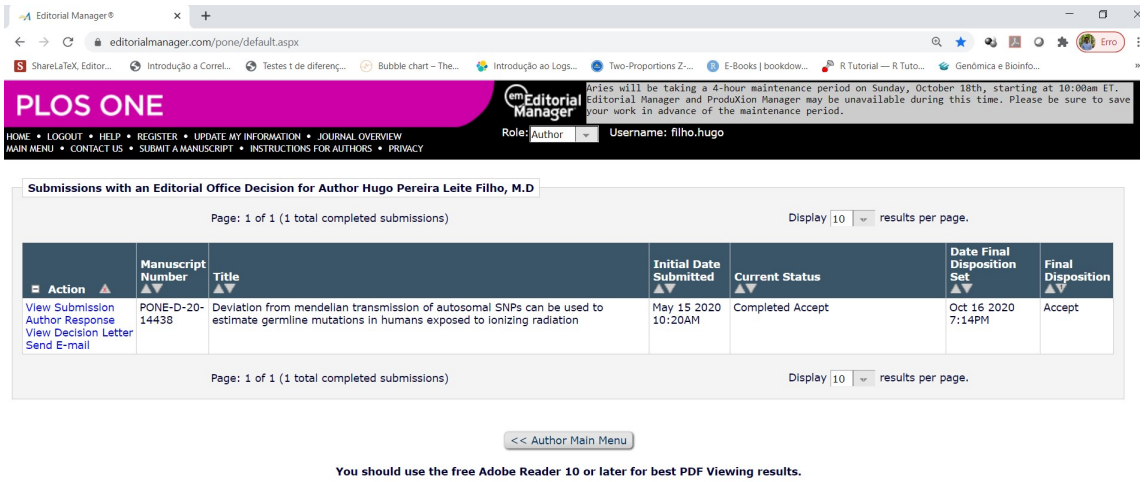
d) Origem da Mutacao: Origem Paterna;

e) Tipo de Substituição: Transicao - Purina/Purina;

f) Tipo de Substituição: Transicao - Purina/Purina;

g) Tipo de Substituição: Tranversão - Purina/Pirimidina

h) Tipo de Substituição: Transicao - Pirimidina/Pirimidina



Editorial Manager®

editorialmanager.com/pone/default.aspx

PLOS ONE

Editorial Manager

Aries will be taking a 4-hour maintenance period on Sunday, October 18th, starting at 10:00am ET. Editorial Manager and ProduXion Manager may be unavailable during this time. Please be sure to save your work in advance of the maintenance period.

HOME • LOGOUT • HELP • REGISTER • UPDATE MY INFORMATION • JOURNAL OVERVIEW
MAIN MENU • CONTACT US • SUBMIT A MANUSCRIPT • INSTRUCTIONS FOR AUTHORS • PRIVACY

Role: Author Username: filho.hugo

Submissions with an Editorial Office Decision for Author Hugo Pereira Leite Filho, M.D

Page: 1 of 1 (1 total completed submissions) Display 10 results per page.

Action	Manuscript Number	Title	Initial Date Submitted	Current Status	Date Final Disposition Set	Final Disposition
View Submission Author Response View Decision Letter Send E-mail	PONE-D-20-14438	Deviation from mendelian transmission of autosomal SNPs can be used to estimate germline mutations in humans exposed to ionizing radiation	May 15 2020 10:20AM	Completed Accept	Oct 16 2020 7:14PM	Accept

Page: 1 of 1 (1 total completed submissions) Display 10 results per page.

<< Author Main Menu

You should use the free Adobe Reader 10 or later for best PDF Viewing results.

1237

Gmail **Hugo Leite Filho <filho.hugo@gmail.com>**

Notification of Formal Acceptance for PONE-D-20-14438R1 - [EMID:69b8f344abb9cbb5]

1 message

PLOS ONE <em@editorialmanager.com> Fri, Oct 16, 2020 at 8:14 PM
Reply-To: PLOS ONE <plosone@plos.org>
To: Hugo Pereira Leite Filho <filho.hugo@gmail.com>

CC: "Emília Oliveira Alves Costa" emiliaoac@yahoo.com.br, "Daniela de Melo e Silva" silvadanielamelo@gmail.com, "Irene Plaza Pinto" iplazapinto@gmail.com, "Claudio Carlos da Silva" dasilvagenetica@gmail.com, "Alexandre Rodrigues Caetano" alexandre.caetano@embrapa.br, "Aparecido Divino da Cruz" acruz@pucgoias.edu.br, "Lorrayne Guimarães Oliveira" lorraynego@gmail.com, "Alex Silva da Cruz" a.silva.cruz@hotmail.com

PONE-D-20-14438R1
Deviation from mendelian transmission of autosomal SNPs can be used to estimate germline mutations in humans exposed to ionizing radiation

Dear Dr. Leite Filho:

I'm pleased to inform you that your manuscript has been deemed suitable for publication in PLOS ONE. Congratulations! Your manuscript is now with our production department.

If your institution or institutions have a press office, please let them know about your upcoming paper now to help maximize its impact. If they'll be preparing press materials, please inform our press team within the next 48 hours. Your manuscript will remain under strict press embargo until 2 pm Eastern Time on the date of publication. For more information please contact onepress@plos.org.

If we can help with anything else, please email us at plosone@plos.org.

Thank you for submitting your work to PLOS ONE and supporting open access.

Kind regards,
PLOS ONE Editorial Office Staff

on behalf of
Dr. Roberto Amendola
Academic Editor
PLOS ONE

In compliance with data protection regulations, you may request that we remove your personal registration details at any time. (Use the following URL: https://www.editorialmanager.com/pone/login.asp?a=r). Please contact the publication office if you have any questions.

1238
1239

Deviation from mendelian transmission of autosomal SNPs can be used to estimate germline mutations in humans exposed to ionizing radiation

Short title: Deviation from mendelian transmission in SNPs of humans exposed to ionizing radiation

Hugo Pereira Leite Filho[1,2], Irene Plaza Pinto[3], Lorraynne Guimarães Oliveira[3,4], Emília Oliveira Alves Costa[3], Alex Silva da Cruz[3], Daniela de Melo e Silva[3,4], Claudio Carlos da Silva[1,2,3,5], Alexandre Rodrigues Caetano[6], Aparecido Divino da Cruz[1,3,4,5,*]

[1] Programa de Pós-Graduação em Biotecnologia e Biodiversidade, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[2] Universidade Estadual de Goiás, Goiás, Brazil
[3] Núcleo de Pesquisa Replicon, Mestrado em Genética, Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás, Goiânia, Goiás, Brazil
[4] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[5] Laboratório de Genética Molecular e Citogenética Humana, Laboratório Estadual de Saúde Pública Dr. Giovanni Cysneiros, Secretaria de Saúde Pública do Estado de Goiás, Goiânia, Goiás, Brazil
[6] Embrapa Recursos Genéticos e Biotecnologia, Brasília, Distrito Federal, Brazil

* Corresponding author: acruz@pucgoias.edu.br (ADC)

Abstract

We aimed to estimate the rate of germline mutations in the offspring of individuals accidentally exposed to Cesium-137 ionizing radiation. The study included two distinct groups: one of cases, consisting of males and females accidentally exposed to low doses of ionizing radiation of Cs^{137}, and a control group of non-exposed participants. The cases included 37 people representing 11 families and 15 children conceived after the accident. Exposed families incurred radiation absorbed doses in the range of 0.2 to 0.5 Gray. The control group included 15 families and 15 children also conceived after 1987 in Goiânia with no history of radiation exposure. DNA samples from peripheral blood were analyzed with the Affymetrix GeneChip® CytoScanHD™ to estimate point mutations in autosomal SNPs. A set of scripts previously developed was used to detect *de novo* mutations by comparing parent and offspring genotypes at the level of each SNP marker. Overall numbers of observed Mendelian deviations were statistically

significant between the exposed and control groups. Our retrospective transgenerational DNA analysis showed a 44.0% increase in the burden of SNP mutations in the offspring of cases when compared to controls, based on the average of MF_{MD} for the two groups. Parent-of-origin and type of nucleotide substitution were also inferred. This proved useful in a retrospective estimation of the rate of *de novo* germline mutations in a human population accidentally exposed to low doses of radiation from Cesium-137. Our results suggested that observed burden of germline mutations identified in offspring was a potentially useful biomarker of effect to estimate parental exposure to low doses of IR and could become an important marker suitable for biomonitoring human population exposed to environmental mutagens.

Introduction

In 1987, a series of unexpected events resulted in a major radiological accident in Goiânia, Goiás, Brazil, causing human, animal, plant and environmental exposure to gamma ray ionizing radiation (IR) of Cesium-137 and contamination by the radionuclide [1]. For some people, individual exposure resulted from internal and external contamination by the radioactive salt, while others were exposed to radiation emitted by the decay of Cesium-137. In some cases, people were both exposed to radiation and contaminated by the radionuclide. In the aftermath, 249 people were exposed to IR from Cesium-137, leading to individual absorbed doses of IR ranging from 0 to 7 Gy, resulting in four fatalities during the acute phase of the accident [2,3].

Following the accident, the exposed population has been extensively monitored using genetic biomarkers, as they have been shown to be efficient biomarkers of exposure to gamma rays [4]. However, each biomarker tends to reveal a distinct biological phenomenon in the exposed cells, mostly associated with DNA repair and how cells physiologically coped to survive a specific insult. In this context, our group

and others have established somatic mutation frequencies using data from glycophorin A [5] and HPRT assays [6], chromosomal aberrations [7,8,9], BCL2/J(H) translocation [10], and micronucleus frequencies [11] in T-lymphocytes of the cohort accidently exposed to Cesium-137 IR. Moreover, in order to understand the effect of IR on the induction of germ line mutations, STR markers [12] were initially used to estimate the rate of germline mutations in the offspring of the exposed cohort. More recently, CNVs have been used as biomarkers for parental exposure to demonstrate the effect of low absorbed doses of IR on germline mutations in the cohort's offspring conceived after the accident [13].

Radiation-absorbed dose relates to the estimated quantity of energy deposited in the mater per unit of mass. Thus, it can be used as an indirect measurement of the harmful biological effect of the radioactive energy on the cellular system. It is calculated by estimating the concentration of energy from radiation exposure deposited in each organ, using a reference value, the type of radiation and the potential for radiation-related mutagenic changes in each organ or tissue [14].

The exposure of cells to IR delays the normal progression of the cell cycle [15-17], initially observed as a passive cellular response resulting from of the induction of DNA damage in the exposed cells. The irradiated cell must adapt to the insult and facilitate DNA repair processes, especially fixing double-strand breaks, the most common damage after DNA exposure to IR [18-22].

The mutagenic effects of IR on the human germ line cells are of concern, as they lead to the accumulation of mutations in the offspring of irradiated parents, amounting to an increase in the mutational burden [23]. Despite numerous studies, little is known about the genetic effects of low doses of radiation from low linear energy transfer gamma radiation exposure in humans. Most of the consolidated evidence comes from

the extrapolation of the induction of germline mutations in mammals, often rat and mouse models [24,25].

Advances in the methodologies of genomic analysis have greatly increased the volume of nucleotide sequence data, enabling the identification of thousands of SNPs (single-nucleotide polymorphisms). Variations in SNPs are important to determine genotypic and phenotypic relationships, within and between species and populations, and also to identify variants related to genetic diseases in humans and animals [26]. In this context, genomic analysis can be a useful tool to study and understand the effects of IR exposure on animals and humans [13,27].

In recent decades, several genotyping technologies have been developed to characterize SNPs all producing genotype matrices with hundreds of thousands of datapoints. Algorithms based on parametric and nonparametric statistical models have been used to determine the genotype of each SNP from the fluorescence signal intensity of marked probes, which are scanned, captured, and arranged in a matrix format [28,29]. One commercially available SNP array, the GeneChip® CytoScanHD™ (Thermo Fisher Scientific, Waltham, MA, USA), is considered to be a high-density matrix, including about 750,000 polymorphic markers with an average genotyping accuracy of >99% [30].

In the aforementioned context, the general objective of the current study was to quantify Mendelian deviations (MD) in genome-wide autosomal SNP data from a cohort of people conceived after parental exposure to Cesium-137 IR, and a group of non-exposed people from the same geographical area. The rate of MD was applied to evaluate if the observed burden of germline mutations identified in the offspring could be a potentially useful biomarker of parental exposure to low doses of IR.

Material and Methods

Sample collection, processing, and genotyping

The experiment was designed as a case-control observational study. The group of cases consisted of 11 families, of whom at least one of the parents was accidentally exposed to IR during the Cesiu-137 accident, totaling 37 participants (11 fathers, 11 mothers, and 15 children conceived after the accident). The radiation absorbed doses for the exposed parents ranged from 0.2 to 0.5 Gy [3,13]. As controls, biological samples were obtained from 15 families living in Goiânia since the time of the accident with no prior history of exposure to IR. Thus, the control group was comprised of 15 fathers, 15 mothers, and 15 children also conceived after 1987. A total of 82 subjects were used in the study whose DNA samples were analyzed using the SNP-array GeneChip® CytoScanHD™ (Thermo Fisher Scientific).

Cases and controls participated voluntarily in the study, which was approved by the ethics committee on research with humans from the Pontifical Catholic University of Goiás (PUC-Goiás) – CAAE number 49338615.2.0000.0037. At the time of blood collection, participants answered a lifestyle questionnaire and signed an informed consent form. A total of 10 mL of peripheral blood in EDTA was voluntarily donated by all participants. Total genomic DNA was isolated from whole blood using Illustra blood genomicPrep Mini Spin Kit® (GE Healthcare, Milwaukee, WI, USA) and stored at −20°C. The remaining biological material was stored according to CNS Resolution 441/11.

Chromosomal microarray analyses were carried out in GeneChip CytoScanHD® arrays (Thermo Fisher Scientific) in order to collect individual genotypes from polymorphic autosomal markers. SNP genotypes were generated using ChAS© (Thermo Fisher Scientific). Every array met the quality controls recommended in the manufacturer's guidelines. SNP genotypes were filtered based on individual call confidence levels for each marker, thus calls with confidence levels $<5x10^{-2}$ and invalid

(no call or null) in one or more samples were removed from the dataset. Therefore, only markers with quality-controlled genotypes in all samples were considered for the analysis. Genotyping was based on the hg19 version of the human genome hosted on the UCSC Genome Browser (University of California, Santa Clara, CA, USA). We also applied the CpG island track from UCS browser in order to stablish the rule out C>T mutations at CpG sites. As the array genotypes didn't allow the discrimination from which strand the damage was derived, all substitutions were included in the data sets.

Principle Component Analysis

Principle Component Analysis (PCA) methods were used to assess whether participants in the case and control groups came from the same genetic population, the dataset contained about 522K SNPs. This step was also included to assess whether individual sample quality effects may have generated spurious results. SNPs with minor allele frequencies (MAF) below 0.01 were removed from the dataset, including all mendelian errors in the samples. Data pruning of the final dataset was performed using the PLINK (2.0) package [31] to generate a subset of markers for PCA analysis using the following parameters: window size of 500 SNP with a step size of 5 SNP, using an r^2 threshold of 0.1. Pruning resulted in a subset of 2.789 SNPs that were used to estimate principal components and to generate plots for each test group.

Analysis and phasing of genotyping data to identify Mendelian deviations

MDs were inferred with a set of previously developed Perl scripts and R libraries [32] termed SIPO (Scripts for Inference of Parental Origin) to mine SNP data in MySql© format. The SIPO pipeline was listed in (Fig 1) and supporting information files are accessible in a GitHub under the accession URL: https://github.com/hugofilho/sipo. Parent genotypes were compared with respective offspring genotypes for each individual marker. Sex chromosome data were excluded

from the analysis, as X-linked data showed elevated noise and Y-specific regions had low marker coverage. Table 1 shows all data variables considered by SIPO.

Table 1. **Variables generated by ChAS® and considered by SIPO to identify MDs**

Marker	SNP marker identifier
Genotyping	Genotyping call. Biallelic information. Three possible combinations: AA/BB/AB
Trust Value	Confidence value for each genotyping call
Sign A	Gross sign value for sign A on the marker
Sign B	Gross sign value for sign B on the marker
Nitrogen Base	Call from the nitrogen base. Biallelic information. Ten possible combinations: AA|AC|AG|AT|CC|CG|CT|GG|TT
dbSNP	SNP identifier record in the NCBI dbSNP database
Chromosome	Autosome associated with the marker
Chromosome Position	SNP locus on the chromosome

First, SIPO validated the .CYCHP file generated by ChAS®, then SIPO identified trio variables and started to generate inferences for *de novo* mutations, corresponding to MDs in the child. Parental origin of observed mendelian deviations were inferred using basic expected mendelian inheritance rules applied over family trio data. For instance, if parent 1 had a genotype "AA" and parent 2 had a genotype "CC", and their child had genotype "GC", the germline mutation was inferred to parent 1. Executed steps allowed to determine nucleotide substitution type in addition to inferring the parent of origin of the MD observed in the offspring. Derived information was loaded into a MySQL database and R scripts were used to perform linear regression, clustering and PCA with the resulting data (Fig 2).

In some situations, SIPO was not able to identify the parental origin of a SNP based on Mendelian transmissions. To solve this challenge, two deductions were incorporated into the pipeline. The first deduction was coded into SIPO to identify confidence interval values of individual SNPs using ChAS® data from the parents of a family trio. The second consisted of identifying the nearest mutated SNP, based on Euclidean distance using tools from Microsoft Excel® (version 365), which had the parent of origin previously inferred following mendelian transmissions rules. Thus, at

the end of the pipeline, the deductions aided to attribute the origin of a mutation to the parent who had both the lowest confidence interval value for that particular mutated SNP and who transmitted that chromosomal segment to the child based on the nearest variant SNP.

The total count of MD was used to estimate the germline mutation frequency (MF_{MD}) in the offspring, using equation 1: [13,33]

$$MF_{MD} = \frac{\sum T_{MD}}{bxnvp} (1)$$

Where ΣT_{MD} = Total MD; b is a biallelic locus (2); *nvp* is the number of valid SNPs in the array according to the assembly of the human reference sequence (GRCh37/hg19) as indicated in S1 Table.

In the present study, all statistical tests were performed considering a 95% confidence interval and 5% significance level. The statistical tests used were the Shapiro-Wilk test, Student's T test, regression analysis, clustering [34], and principal component analysis [35,36]. The R statistical package [32] was used in all analyses.

Results and discussion

The current study used SNP genotypes from a cohort of offspring born to parents accidentally exposed to Cs-137 to estimate the induction of germline mutations in humans exposed to low doses of ionizing radiation. As a cautionary note, in the current work, deviation of a Mendelian transmission implies that a point mutation observed in a child wasn't observed in his/her parents, thus it was herein interpreted as a *de novo* mutation. However, we are aware that SNP variants can rise somatically due to DNA repair failure in the first cell divisions of the embryo, a variable common to both cases and controls and expected to be equally represented in the study datasets, bearing little bias to the dataset if any.

Before disclosing the results of the study, we also wish to note the limitation regarding the small size of the study cohort, which could render meaningful conclusions at first glance. In this context, two important rationales support the value of considering follow-up studies of human populations exposed to IR. First, considering the global effort with respect to radioprotection and regulation, it's very unlikely that large accidentally exposed cohorts will be available world-wide to be investigated with the newest methodologies. Second, a high-density SNP array was used to call thousands of SNPs, covering a very large proportion of the genome. Thus, increasing the chances of identifying genomic variation in small populations that could be potentially useful to establish new biomarkers of effect to be applied in future studies investigating genotoxic and mutagenic responses to environmental stressors. The current available technologies applied to the study of genomes have that intrinsic characteristic, making them tools of first-tier choice in a variety of investigations, particularly when assessing small cohorts.

CytoScan HD Suite had an intrinsic algorithm, which allowed the analysis of a chromosome segment given the presence of polymorphic markers within that region. In

the current study, the challenge was to establish the parent-of-origin for a point mutation based solely on Mendelian transmissions. In order to infer that origin, two deductions were incorporate into our pipeline, which allowed the inclusion of 9,522 and 4,821 MDs for case and control groups, respectively, into the dataset. In this context, the current pipeline could be used as an additional tool to define the parental origin of polymorphic variants obtained from SNP array genotypes.

PCA results using a subset of LD-pruned data (522 Kb SNPs) indicated subjects included in both case and control groups belonged to the same population and there were no recognizable additional confounding factors associated with the test groups other than exposure to Cs-137 (Fig 3). Therefore, the MF_{MD} could be compared between groups, even with a reduced sample size. Observed MDs followed a normal distribution ($p = 0.5592$) and were all included in subsequent statistical analyses. The lowest individual numbers of MDs were 972 and 682, and the highest were 2,875 and 1,635 for the case and control groups, respectively (Table 2). Observed MDs were randomly distributed on the SNPs in the array. When performing family trio comparisons, most MDs (60%) were observed only once with no repetition, while 27%, 9% and 4% of the same MDs were respectively observed twice, three and four times in the family trios, confirming both the random effect of DNA damage induced by IR and spontaneous replication errors. Moreover, this observation also favors the quality of the array avoiding artefactual genotyping errors to be included in the dataset.

Table 2. Overall data from both control and exposed groups regarding the study of germline mutation in the offspring of people accidentally exposed to low absorbed doses of Cesium-137 ionizing radiation in Goiania (Brazil).

Group	Family	Exposed Progenitor	Absorbed Dose (Gy)	Paternal Age[2,*]	Maternal Age[2]	Age of offspring	Sex of offspring	MDs[1]				Total of valid SNPs	Frequency of MDs
								Father	Mather	Unknown	Total		
Control	Ct001	None	0	40	36	9	Female	783	694	10	1487	702,304	1,06E-03
	Ct25	None	0	47	36	9	Male	545	631	3	1179	683,381	8,62E-04
	Ct27	None	0	26	26	23	Male	594	729	9	1332	692,311	9,62E-04
	Ct39	None	0	24	24	3	Female	679	679	11	1369	674,320	1,02E-03
	Ct40	None	0	45	37	3	Male	710	711	30	1451	696,544	1,04E-03
	Ct45	None	0	37	31	15	Male	643	663	8	1314	683,243	9,62E-04
	Ct51	None	0	35	34	2	Female	384	479	8	871	697,737	6,24E-04
	Ct52	None	0	55	41	1	Male	1015	588	32	1635	710,477	1,15E-03
	Ct53	None	0	35	27	8	Male	315	361	6	682	713,372	4,78E-04
	Ct60	None	0	40	38	1	Female	501	482	6	989	712,261	6,94E-04
	Ct66	None	0	31	20	26	Female	543	594	2	1139	707,798	8,04E-04
	Ct68	None	0	33	20	10	Male	758	654	11	1423	712,191	1,00E-03
	Ct70	None	0	20	24	8	Female	448	545	3	996	714,892	6,96E-04
	Ct72	None	0	31	20	14	Female	521	614	4	1139	710,259	8,02E-04
	CtF09	None	0	19	21	25	Male	718	696	9	1423	712,352	9,98E-04
Exposed	Ex04	Father	0.1	27	27	20	Male	900	966	7	1873	698,305	1,34E-03
	Ex06	Father	0.3	35	26	9	Male	1045	1082	9	2136	693,858	1,54E-03
	Ex07-1F	Mother	0.2	54	24	19	Male	976	850	8	1834	692,375	1,32E-03
	Ex07-4F	Mother	0.2	56	26	17	Female	1509	1085	18	2612	689,467	1,89E-03
	Ex08	Mother	0.2	18	20	8	Male	538	718	4	1260	706,759	8,92E-04
	Ex10	Mother	0.2	21	24	2	Female	1187	1054	21	2262	705,993	1,60E-03
	Ex12	Mother	0.3	31	30	3	Male	1361	1486	28	2875	693,463	2,08E-03
	Ex15	Father	0.2	18	27	16	Male	560	645	7	1212	706,780	8,58E-04
	Ex18	Father	0.2	47	30	18	Male	819	800	12	1631	707,366	1,15E-03
	Ex21	Mother	0.2	38	27	20	Female	457	513	2	972	707,827	6,86E-04
	Ex22-2F	Mother	0.2	29	31	20	Female	1006	664	3	1673	707,075	1,18E-03
	Ex22-3F	Mother	0.2	32	34	17	Female	845	566	3	1414	708,278	9,98E-04
	Ex22-4F	Mother	0.2	33	35	16	Male	781	518	5	1304	708,885	9,20E-04
	Ex24	Father	0.5	21	19	12	Female	1010	1102	47	2159	703,635	1,53E-03
	Ex25	Father	0.5	18	16	15	Female	598	708	10	1316	706,598	9,32E-04

[1]Medelian deviations; [2]Age at conception; *All ages are in years old.

1489

In the current study, mutation burden was defined by the number of *de novo* base substitutions in an assayed SNP of a child born to a parent exposed to IR. Thus, a total of 18,429 and 26,533 SNPs showed MD for control and cases, respectively. Thus, the overall frequencies of germline mutations observed in the different trios were, on average, 1.3 $x10^{-3}$ and 0.9 $x10^{-3}$ mutations per polymorphic marker. The Student's T test showed the difference in the means was statistically significant assuming equal variances for both groups (p=0.002). Tables 2 and 3 contain the summary of the data used in this study. Our retrospective transgenerational DNA analysis showed about a 44.0% increase in the burden of SNP mutations in the offspring of cases when compared to controls, based on the average of MF_{MD} for the two groups. The current study pioneered the application of SNP data analysis to identify MD and estimate germline mutations in the offspring of humans accidentally exposed to low absorbed doses of IR. Current findings corroborated our first study reporting the usefulness of small CNVs to estimate *de novo* human germline mutation rates in a similar cohort [13]. A previous study by [23] also described the usefulness of the mutation frequencies of *de novo* CNV and SNVs as biomarkers of effect for paternal exposure to IR in mice. Moreover, a recent study using whole genome sequencing data from an offspring of radar soldiers potentially exposed to IR found the differences in the frequency of *de novo* SNVs might be suited for the assessment of DNA damage from IR in humans [37].

Table 3. Summary of the descriptive data of the case and control groups for parental and F1 generations in the study of the effect of IR exposure on the induction of germline mutations in humans.

Generation	Variables		Cases	Control
Parental	N		15	15
	Age range (years)		16 − 56	19 – 55
	Mean age at conception (years ± SD*)	Paternal	31.9 (12.5)	34.4 (9.9)
		Maternal	26.4 (5.3)	28.9 (7.4)
	Absorbed dose (Gy)		0.2 − 0.5	0
F1	N		15	15

Age range (years)	2 – 20	1 – 26
Mean age (years ± SD*)	14.0 (5.9)	10.5 (8.5)
Sex ratio (Male:Female)	8:7	8:7
Mean MF_{MD} (± SD*)	1.3×10^{-3} ($\pm 0.4 \times 10^{-3}$)	0.9×10^{-3} ($\pm 0.2 \times 10^{-3}$)

*SD = Standard Deviation.

We also carried out a liner regression in order to evaluate the relationship between the radiation-absorbed doses and the MF_{MD} in our cohorts. Our results were statistically significant (p=0.004; R^2=0,257), suggesting that low absorbed doses of IR could predict an increase of the mendelian deviation in the exposed group, which could be linearly fitted (Fig 4) following the equation below:

$$MF_{MD} = 0.001 + 0.001(dose)$$

To date, there is extensive evidence supporting sex differences in mutation rates, with older fertile males expected to contribute more to the burden of a mutational health hazard than older females. A greater number of continuous cell divisions in the male germ line has been implicated as one reasonable explanation for such difference on paternal age effect [38,39]. However, although this have been consistently reported, a clear and definite conclusion on the subject remains to be reached [38,39]. In our study, the sex of the progenitors had no effect on the MF_{MD} of autosomal SNPs as for both case and control groups mothers and fathers contribute equal numbers of *de novo* MD to their offspring. When taken into consideration the sex of the exposed parent, the average of the frequencies of germline mutations of children born to exposed fathers was 1.2×10^{-3} ($\pm 0.3 \times 10^{-3}$) and for exposed mothers was 1.3×10^{-3} ($\pm 0.5 \times 10^{-3}$), with no statistical differences (p=0.195) intragroup.

With respect to the potential parental age effect, our control group revealed older fathers contributed more MDs to their offspring (Figs 5A-5C) , which could be modeled by the number of mitotic spermatogonia divisions as a function of age, reinforcing previous findings regarded as male-mutation bias [39,40]. However, our study failed to

detect the maternal age effect on the number of MDs (Fig 6). Although there has been increasing evidence of maternal contributions to the *de novo* point mutations in the offspring [41,42], others have argued that females contribute less MD to their offspring based on sex differences in gametogenesis and development [43]. To date, there is an ongoing debate about the maternal and paternal contributions to the germline mutation burden in the offspring [44]. New genomic and statistic tools applied to large and diverse populational datasets will soon help bring forth a resolution for this biological conundrum. Although larger number of family trios might be needed to assess the female contribution on the germline point mutations in their offspring, our results suggested that strength of male-mutation bias could be observed even in small family cohorts.

Single base substitutions have been a common and frequent mutational event subjacent to cell divisions spontaneously that rise as a consequence of DNA replication errors or induced by environmental stressors, such as IR. Some previously published studies on the types of DNA spontaneous base substitutions indicated all possible substitutions are well represented in germline cells [45]. Such studies suggested that transition rates tend to be higher [46] than transversion rates [47]. The findings in the current study supported these observations, since a higher proportion of transitions was observed in the children from both cases and controls.

It has been generally assumed that in groups of small sample sizes, it would be very difficult to detect the maternal age effect on the burden of point mutations in the offspring. Nevertheless, in order to test the hypothesis that in our exposed cohort germline mutations in both sexes were damage-induced by exposure to low doses of IR, we stratified our set of phased *de novo* mutations in 6 classes based on parental and derived alleles (Table 4).

Table 4. Summary of the descriptive data of the case and control groups for the six classes of base substitution in the genome of children conceived after parental exposure to low doses of ionizing radiation and their controls.

lGroup	Class	Minimum	Maximum	Mean	SD*	Total
Control	C>A	38	137	100.93	29.058	1,514
	C>G	45	141	111.47	28.538	3,344
	C>T	238	558	435.93	92.669	2,798
	T>A	45	84	64.67	14.034	6,539
	T>C	252	635	418.93	99.427	970
	T>G	61	141	96.67	21.091	6,284
Exposed	C>A	72	209	148.27	45.325	2,224
	C>G	90	291	168.60	56.616	2,529
	C>T	376	1.166	672.67	234.248	10,090
	T>A	48	147	85.27	27.825	1279
	T>C	307	847	561.60	161.834	8,424
	T>G	75	219	132.47	42.797	1,987

*SD=Standard Deviation

All the SNPs harboring C>T transitions in the data sets were not located in CpG islands and were all included in the analyses. IR is known to cause double stand brakes and all types of base substitutions. Although all transitions and transversions were observed in our data set (Fig 7), C>T and T>C were overrepresented, for both cases and controls, favoring the well-known hypothesis that human genome harbor a mutational bias toward A/T composition in the DNA stand [48]. In our study, although the base line of the MF_{MD} in SNPs were different, the mutational spectra of cases and controls, considering all base substitutions, were remarkably similar. This observation supports previous claims regarding the random effect of the deposition of radiation energy on biological systems [49].

In the context described before, MF_{MD} of polymorphic markers was a quantifiable and useful variable to estimate the parental contribution to the mutational burden of their children, as a consequence of transmitting non-deleterious point mutations induced by IR above the threshold expected from the control population. DNA damage in the parental germ lines could have gone uncorrected by the DNA repair

system, fixed in the cells and then transmitted to the offspring. The F test, to evaluate MD frequencies in the test groups, showed the number of observed MDs were significantly different ($F = 4.47$; $p = 8 \times 10^{-3}$). The arithmetic mean of the MD in the offspring of case and control groups are shown in Fig 8A, whereas Fig 8B shows the representation of the total of MDs observed in each family trio in both groups.

To validate the findings of the current study, which analyzed the MF_{MD} of a small cohort of children conceived after their parents were accidentally exposed to ionizing radiation from Cs-137, we suggest the application of the current study design to larger cohorts. It might be advisable to include a wider range of absorbed doses, resulting from either therapeutic or occupational exposures, to assess the potential of Mendelian deviations as retrospective biomarkers for IR exposure in human populations. In the present study, the case and control groups belonged to the same population and, therefore, were subjected to similar general environmental effects. Thus, it was safe to conclude that the average MF_{MD} was higher in the exposed group as a result of higher germline base substitutions than in the control group, which could be reasonably assumed as a consequence of parental exposure to low doses of IR. In this context, low doses of low-LET radiation induced MD in autosomal SNPs that could be identified, quantified and, therefore, used as a biomarker of effect to study human populations according to their history of exposure to environmental mutagenic insults.

Conclusions

This study pioneered the analysis of MDs using autosomal SNP data observed in parent-offspring trios as biomarkers of effect to low doses of ionizing radiation. We succeeded estimating retrospectively the germline mutation frequency of SNPs in a human population accidentally exposed to low doses of radiation from Cs-137 and estimated the burden of germline mutations in the offspring.

We found the sex of the progenitors had no effect on the MF_{MD} of autosomal SNPs, for both case and control groups, mother and fathers contributed equal numbers of *de novo* MD to their offspring. After accounting for age, our control group revealed older fathers contributed more MD to their offspring, which could be modeled by number of mitotic spermatogonia divisions as a function of age, supporting previous findings of male-mutation bias. However, our study failed to detect the maternal age effect on the frequency of MDs.

In summary, there was a 44.0% increase in the MF_{MD} of the offspring of those accidentally exposed to low doses of IR, from a radiological accident in Goiânia. Low absorbed doses of IR could predict the increase of the mendelian deviation in the exposed group. Therefore, we concluded that MF_{MD} is a potentially useful biomarker to estimate parental exposure to IR and suitable for human population biomonitoring. In this context, future studies involving the behavior of MDs following diverse genomic and mutagenic hazards, caused by exposure to environmental agents, may provide important knowledge of the biological effects, mechanisms, and risks resulting from human exposure to such agents.

Finally, we are confident SNP array data can be used to estimate ionizing radiation-induced mutagenesis in human populations, provided the appropriate bioinformatics and statistical tools are used to extract the necessary information for biological inferences and to validate the scientific hypotheses underlying each investigation.

Acknowledgements

At first, the authors wish to express their gratitude to the volunteers who selfishly agreed to participate in the study. We also thank CARA (Centro de Assistência ao Radioacidentado da SES-GO) for helping with contacting the group of exposed parents.

Moreover, we thank Dr. Fernando Nodari and his team for assisting with the issues regarding the commercial genotyping array. Lastly, we would like to thank Mr. Sean Quail for proofreading the manuscript. A.D.C., A.R.C., and D.M.S are CNPq research fellows.

References

1. da Cruz ASP, de Melo e Silva D, Godoy FR, de Melo AV, Costa EOA, Pedrosa ER, et al. Analysis of microsatellite markers located on the Y-chromosome (Y-STR) of individuals exposed to Cesium-137. Studies Goiânia. 2010; 37:799-809.
2. Skandalis A, da Cruz AD, Curry J, Nohturfft A, Curado MP, Glickman BW. Molecular analysis of t-lymphocyte hprt- mutations in individuals exposed to ionizing radiation in goiânia, brazil. Environmental and Molecular Mutagenesis. New York. 1997; 29(2): 107-116.
3. International Atomic Energy Agency (IAEA). The Radiological Accident in Goiânia. IAEA, Vienna. 1988. pp. 1-157.
4. Rana S, Kumar R, Sultana S, Sharma RK. Radiation-induced biomarkers for the detection and assessment of absorbed radiation doses. J Pharm Bioallied Sci. 2010; 2(3): 189–196.
5. Straume T, Langlois RG, Lucas J, Jensen RH, Bigbee WL, Ramalho AT, et al. Novel biodosimetry methods applied to victims of the Goiânia accident. Health Phys. 1991; 60: 71–76.
6. International Atomic Energy Agency (IAEA). Significance and Impact of Nuclear Research in Developing Countries. IAEA, Vienna. 1987; 1:3-14
7. International Atomic Energy Agency (IAEA). The Radiological Accident in Goiânia. IAEA, Vienna. 1988; 1:1-157
8. da Cruz AD, McArthur AG, Silva CC, Curado MP, Glickman BW Human micronucleus counts are correlated with age, smoking, and cesium-137 dose in the Goiânia (Brazil) radiological accident. Mutat Res. 1994; 313: 57–68.
9. da Cruz AD, Curry J, Curado MP, Glickman BW. Monitoring hprt Mutant Frequency Over Time in T-Lymphocytes of People Accidentally Exposed to High Doses of Ionizing Radiation. Environ Mol Mutagen. 1996; 27:165–175.
10. da Silva CC, da Cruz AD. An easy procedure for cytogenetic analysis of aged chromosome preparations using FISH–WCP probes. Chromosome Res. 2002; 10: 233–238.
11. Nunes HF, Laranjeira ABA, Yunes JA, Costa EOA, Melo COA, e Silva DM, et al. Assessment of BCL2/J(H) translocation in healthy individuals exposed to low-level radiation of 137CsCl in Goiânia, Goiás, Brazil. Genet. Mol. Res. 2013; 12(1): 28-36.
12. da Cruz AD, e Silva DDM, da Silva CC, Nelson RJ, Ribeiro L M, Pedrosa ER, et al. Microsatellite mutations in the offspring of irradiated parents 19 years after the Cesium-137 accident. Mutat Res. 2008; 652: 175–179.
13. Costa EOA, Pinto IP, Gonçalves MW, da Silva JF, Oliveira LG, da Cruz AS, et al. Small de novo CNVs as biomarkers of parental exposure to low doses of ionizing radiation of caesium-137. Sci Rep. 2018; 8: 5914.
14. Fazel R, Krumholz HM, Wang Y, Ross JS, Chen J, Ting HH, et al. Exposure to low-dose ionizing radiation from medical imaging procedures. N Engl J Med. 2009; 361(9): 849-857.
15. Iliakis G, Wang Y, Guan J, Wang H. DNA damage checkpoint control in cells exposed to ionizing radiation. Oncogene 2003; 22(37): 5834-5847.
16. Bernhard EJ, Maity A, Muschel RJ, McKenna WG. Effects of ionizing radiation on cell cycle progression. Radiat Environ Biophys. 1995; 34(2): 79-83.
17. Maity A, Mckenna WG, Muschel RJ. The molecular basis for cell cycle

delays following ionizing radiation. Radiother Oncol. 1994; 31(1): 1-13.
18. Mettler FA, Thomadsen BR, Bhargavan M, Gilley DB, Gray JE, Lipoti JA. Medical radiation exposure in the U.S. in 2006: preliminary results. Health Phys. 2008; 95(5): 502-507.
19. Kao GD, Jiang Z, Fernandes AM, Gupta AK, Maity A. Inhibition of phosphatidylinositol-3-OH kinase/Akt signaling impairs DNA repair in glioblastoma cells following ionizing radiation. J Biol Chem. 2007; 282(29):21206-12.
20. Lücke-Huhle C, Comper W, Hieber L, Pech M. Comparative study of G2 delay and survival after ^{241}Americium-α and ^{60}Cobalt-γ irradiation. Radiat Res. 1982; 20: 298-308.
21. Tobey RA. Different drugs arrest cells at a number of distinct stages in G_2. Nature 1975; 254(5497): 245-247.
22. Walters RA, Gurley LR, Tobey RA. The metabolism of histone fractions: Phosphorylation and synthesis of histones in late G_1-arrest. Biophys J. 1974; 164(2): 99-118.
23. Adewoye AB, Lindsay SJ, Dubrova YE, Hurles ME. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. Nat Commun. 2015; 6: 6684.
24. Nakamura N, Suyama A, Noda A, Kodama Y. Radiation effects on human heredity. Annu Rev Genet. 2013;47:33-50.
25. UNSCEAR. Hereditary Effects of Radiation United Nations. United Nations, New York, 2001; 1: 10-15.
26. Shah SC, Kusiak A. Data mining and genetic algorithm based gene/SNP selection. Artif Intell Med. 2004; 3:183-96.
27. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011; 27(21): 3070-1 .
28. Xu Y, Peng B, Fu Y, Amos CI. Genome-wide algorithm for detecting CNV associations with diseases. BMC Bioinform. 2011; 12:331.
29. Dalma-Weiszhausz D, Warrington J, Tanimoto EY, Miyada CG. The Affymetrix GeneChip® platform: An overview. DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols. 1st ed. Academic Press; 2006; 410: 3-23.
30. Zahir FR, Marra MA. Use of Affymetrix arrays in the diagnosis of gene copy-number variation. Curr Protoc Hum Genet. 2015; 85: 8.13.1-8.13.13.
31. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet.;81(3):559-75.
32. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2008.
33. Da Cruz DA, de Melo e Silva D, da Silva CC, Nelson RJ, Ribeiro LM, Pedrosa ER, et al. Microsatellite mutations in the offspring of irradiated parents 19 years after the Cesium-137 accident. Mutat Res. 2008; 652(2):175-9.
34. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster analysis basics and extensions. R package version 2.1.0. 2019
35. Kassambra A. Practical guide to principal component methods in R. Sthda.com, 2017.
36. Husson F, Le S, Pages J. Exploratory multivariate analysis by example using

R, Chapman and Hall. 2010.
37. Holtgrewe M, Knaus A, Hildebrand G, Pantel J. T., de los Santos MRN, et al. Multisite de novo mutations in human offspring after paternal exposure to ionizing radiation. Scientific reports. 2018; 8(1): 1-5.
38. Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Proceedings of the National Academy of Sciences. 2019; 116(19): 9491-9500.
39. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E., et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017; 549: 519–522.
40. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends in Genetics. 2013; 29(10): 575-584.
41. Goldmann JM, Wong WS, Pinelli M, Farrah T, Bodian D, Stittrich AB, et al. Parent-of-origin-specific signatures of de novo mutations. Nature genetics. 2016; 48(8): 935.
42. Wong WS, Solomon, BD, Bodian, DL, Kothiyal, P, Eley, G, Huddleston, KC, et al. New observations on maternal age effect on germline de novo mutations. Nature communications. 2016; 7(1): 1-10.
43. Gao Z, Wyman, MJ, Sella, G, Przeworski, M. Interpreting the dependence of mutation rates on age and time. PLoS biology. 2016; 14(1): e1002355.
44. Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annual review of genomics and human genetics. 2014; 15(1): 47-70.
45. Brovarets OO, Hovorun DM. Proton tunneling in the A·T Watson-Crick DNA base pair: myth or reality? J Biomol Struct Dyn. 2015; 33(12): 2716-2720.
46. da Cruz AD, Glickman BW. Nature of mutation in the human hprt gene following in vivo exposure to ionizing radiation of Cesium-137. environmental and molecular mutagenesis, Environ Mol Mutagen. 1997;30(4):385-95.
47. Lyons DM, Lauring AS. Evidence for the selective basis of transition-to-transversion substitution bias in two RNA viruses. Mol Biol Evol. 2017; 34(12): 3205-3215.
48. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences. 2010; 107(3): 961-968.
49. Vértes A, Nagy S, Klencsár Z, Lovas RG, Rösch F. Dosimetry and Biological Effects of Ionizing Radiation. In: Handbook of Nuclear Chemistry. Boston, MA: Springer; 2003. pp. 1647-1684.

1640
1641
1642

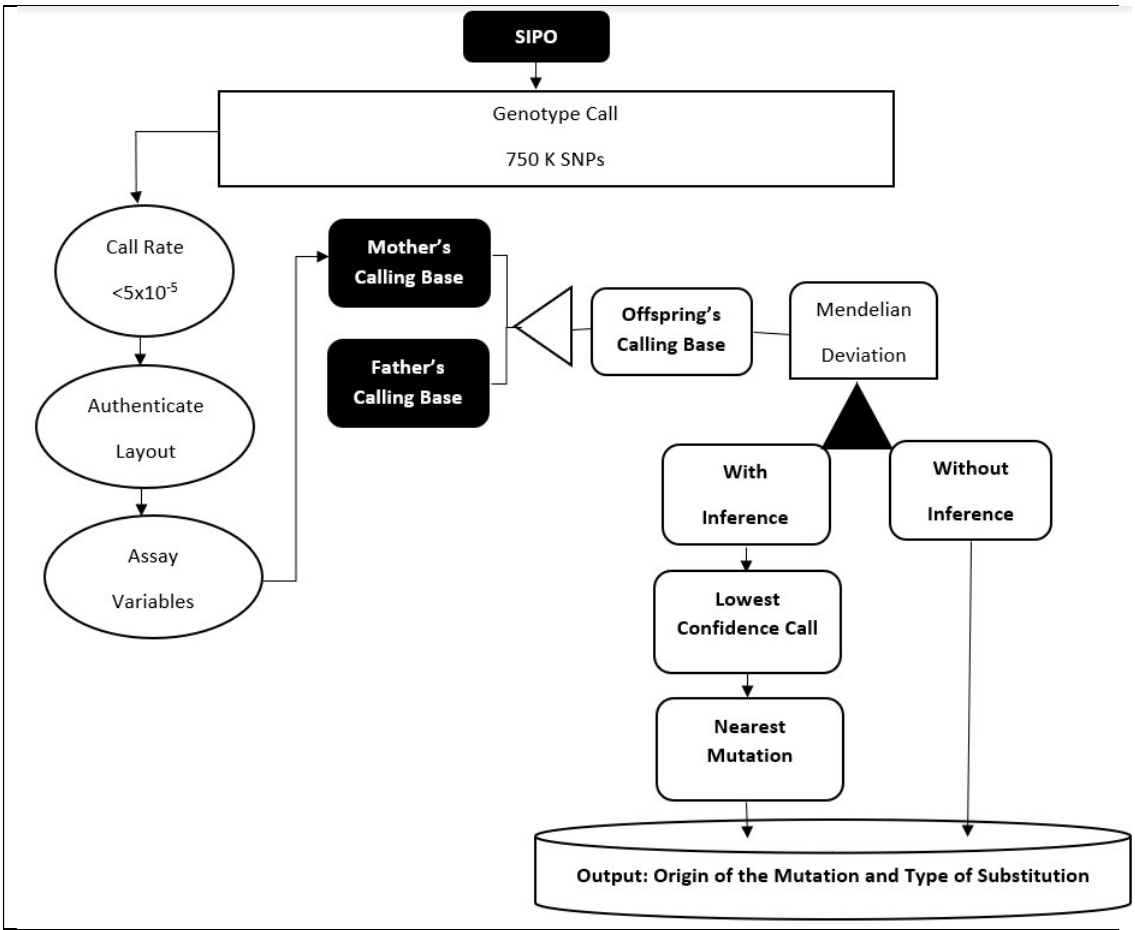


Fig 1. Steps performed by SIPO to infer de novo mutations. Deducting the parental origin of the MD, indicating the type of substitution and generating the estimated rate of Mendelian deviation in the offspring of people accidentally exposed to Cs-137 ionizing radiation.

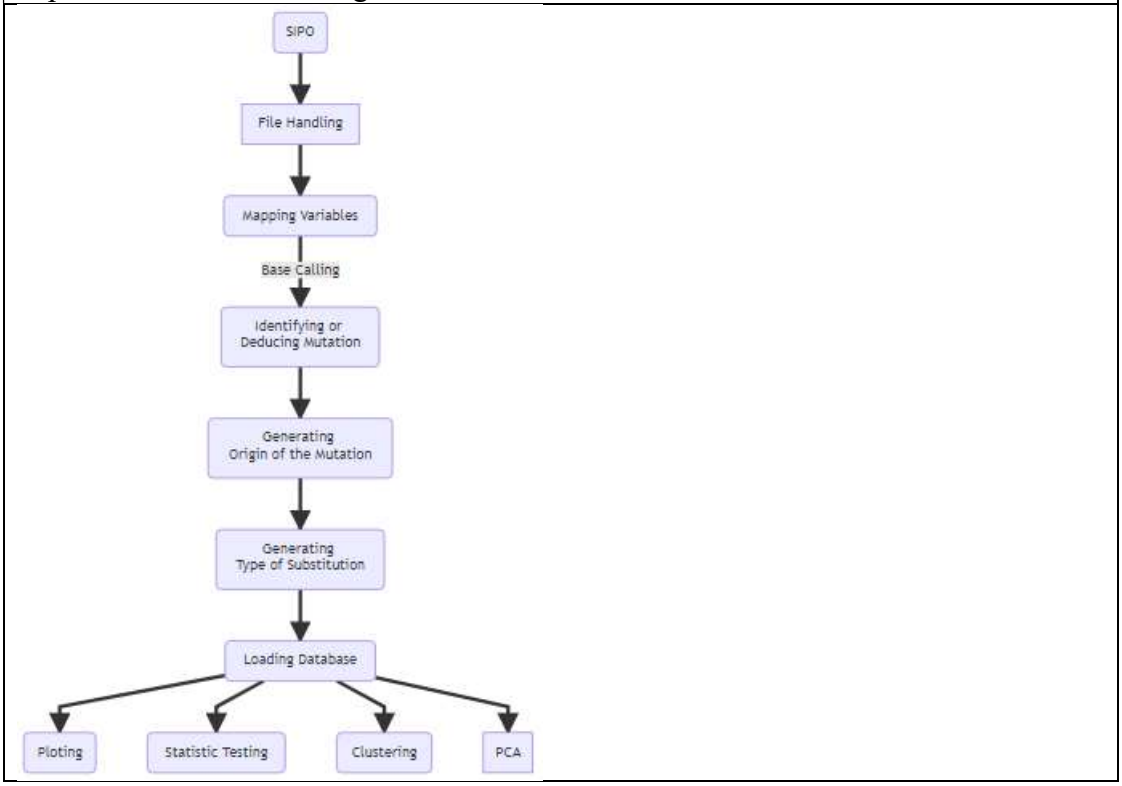


Fig 2. Workflow of the SIPO. Steps performed to infer de novo mutations from SNPs obtained from a high-density genotyping array.

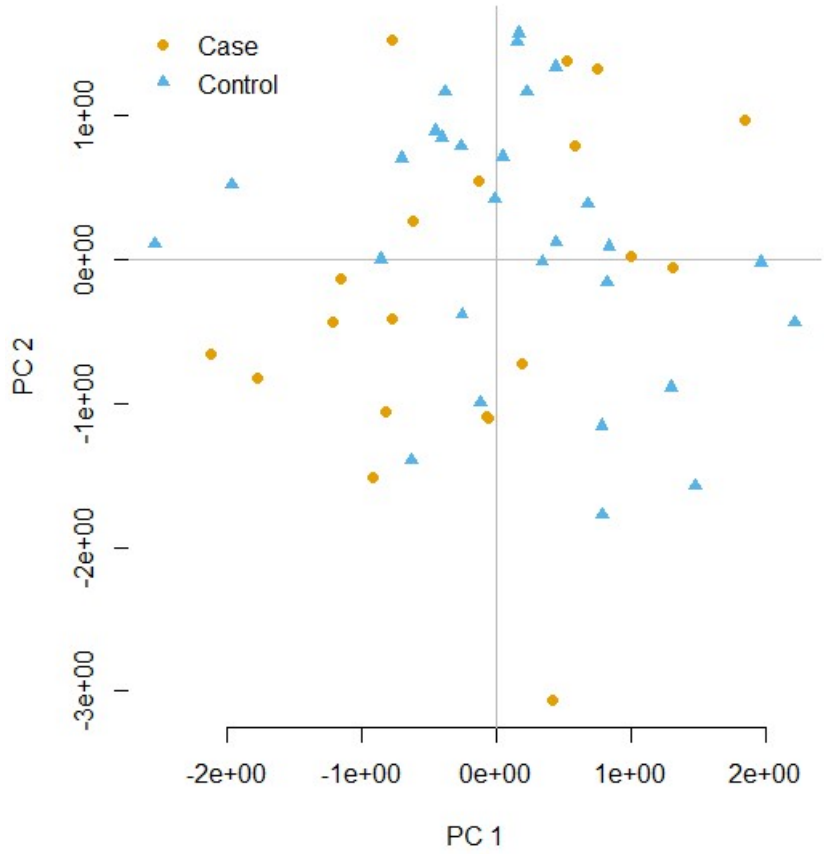


Fig 3. PCA with 2.7K SNP. The variables contained in the PCA represented the standardized relationship matrix by variance, multidimensional sizing (MDS) based on Hamming distances.

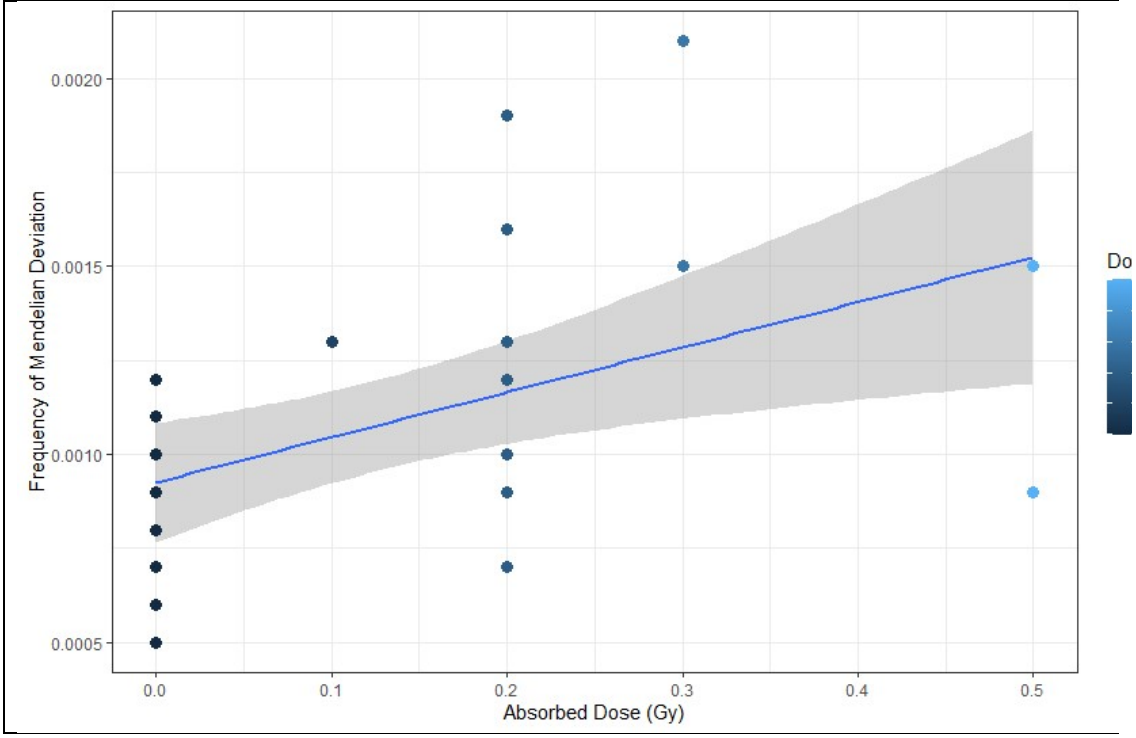
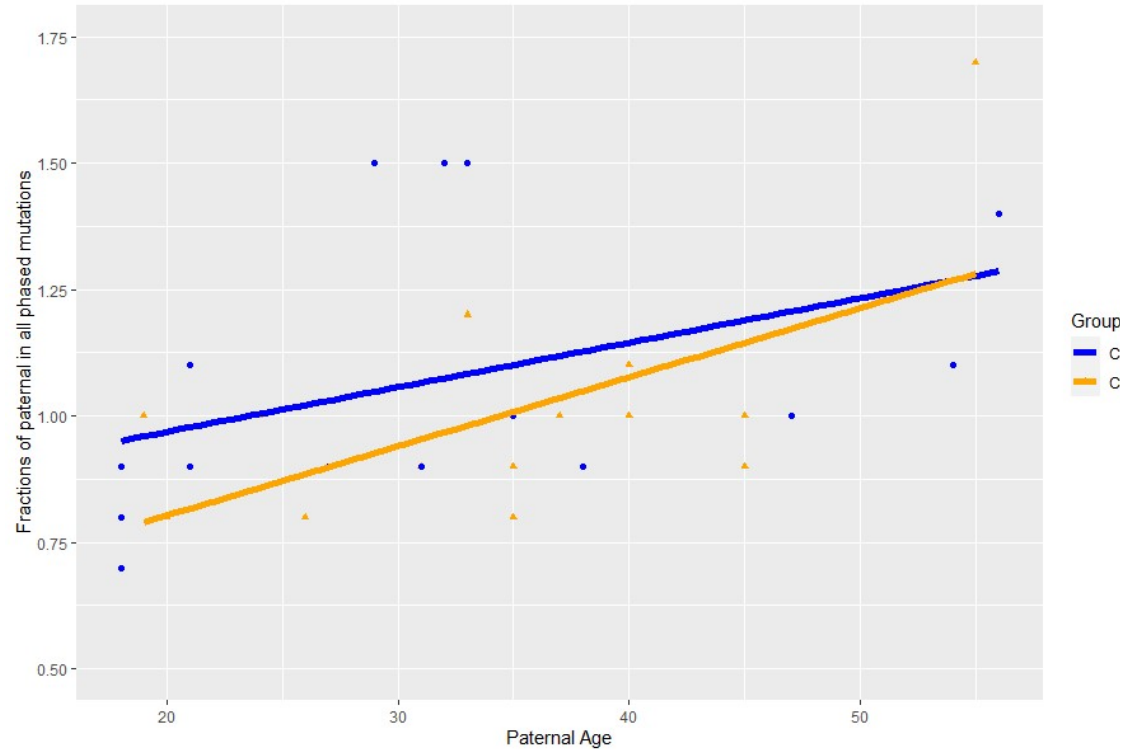
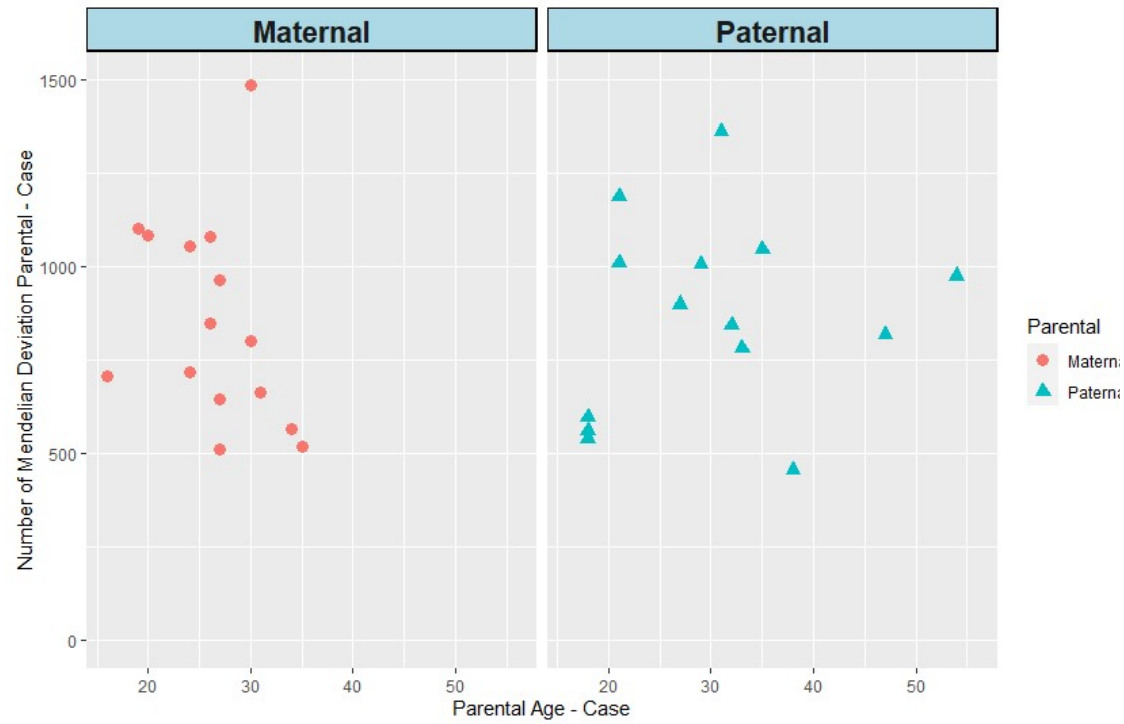


Fig 4. Representation the relationship between the radiation-absorbed doses and the means frequency of Mendelian deviations in a cohort of people conceived after parental exposure to ionizing radiation.

A)



B)



C)

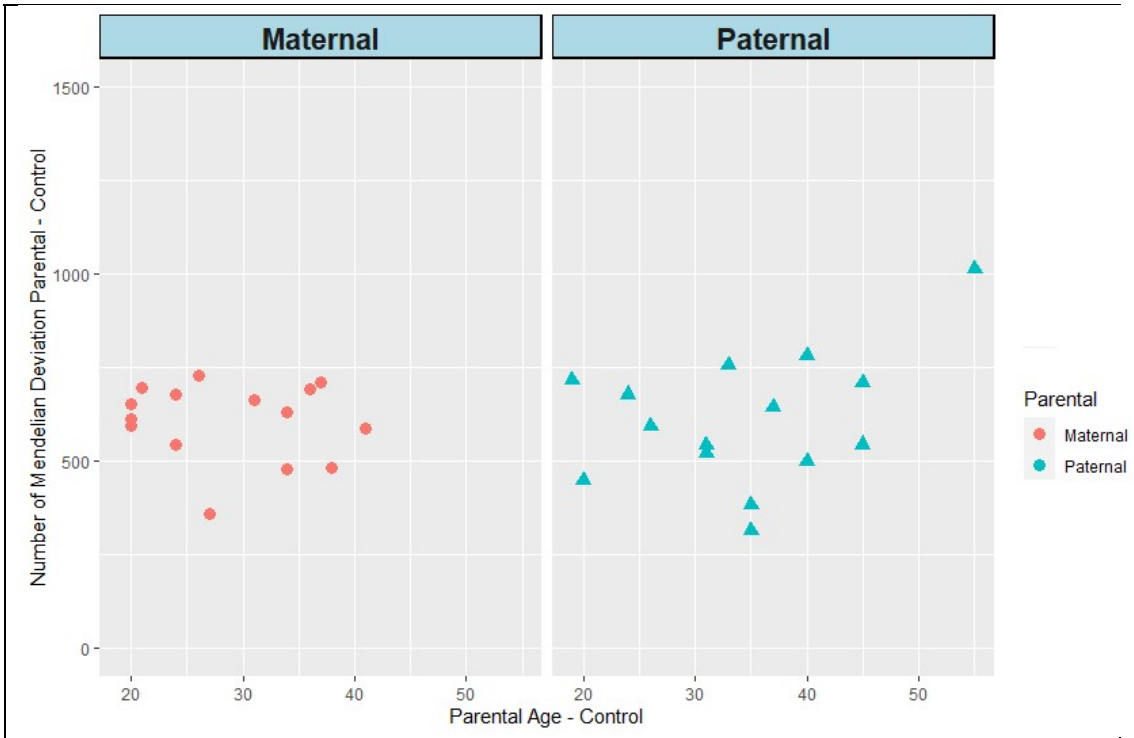


Fig 5. Potential parental age effect. (A) The fraction of paternal mutations as a function of paternal age at conception. Each point represents the data for one child (proband) with similar parental ages. The x-axis position is parental age, the y-axis position is fraction among paternal and maternal. Show the broadcast of data regarding the parental origin that transmitted the mutation to the child in the case group (B) and control group (C).

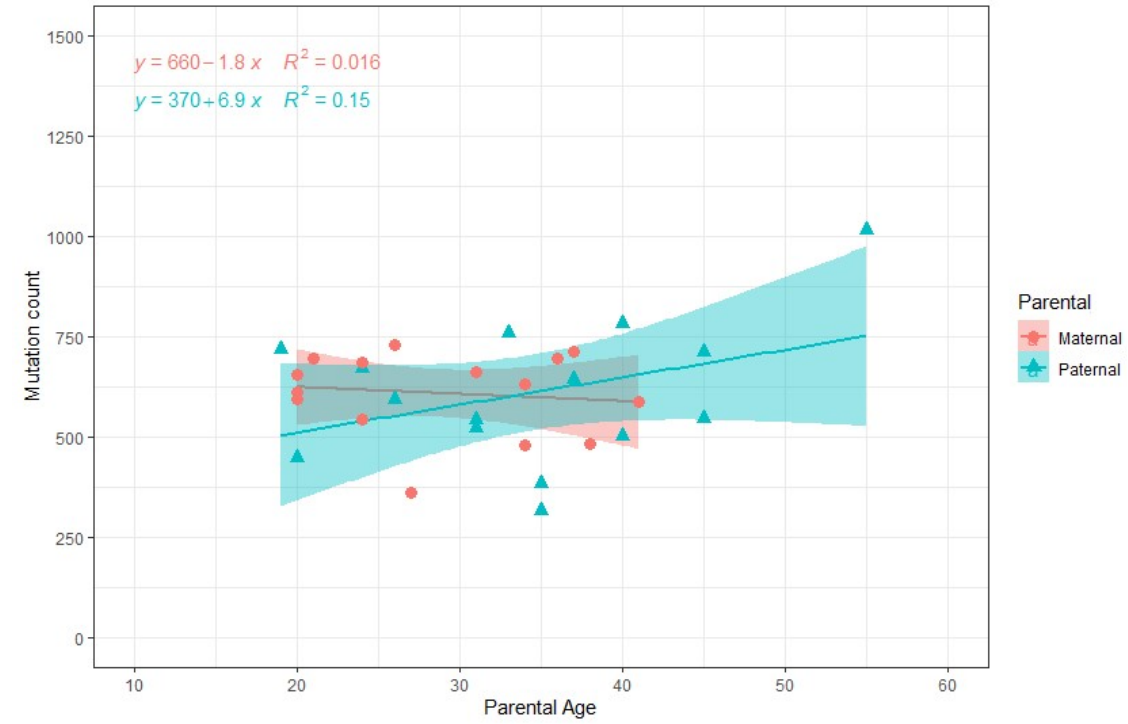


Fig 6. Representation maternal age effect on the number of Mendelian deviations.

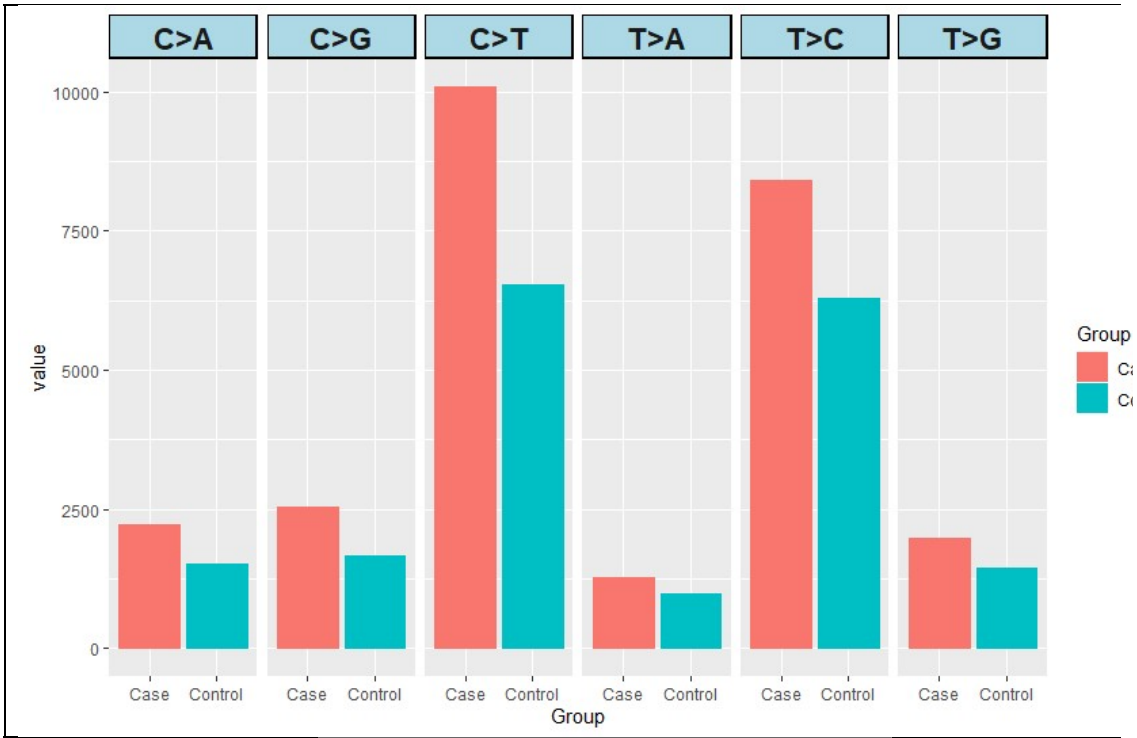
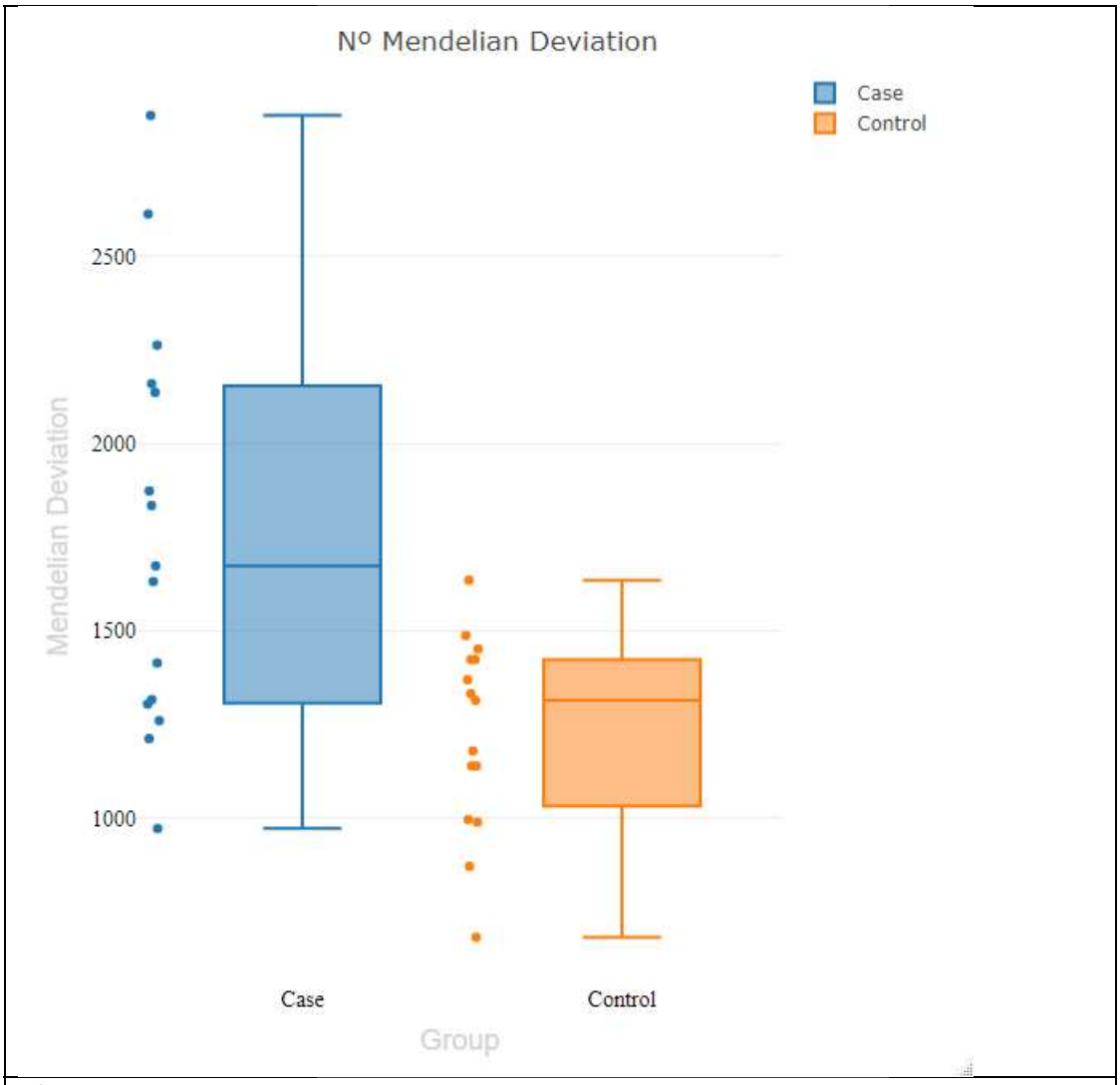


Fig 7. Phased de novo mutations based on parental and derived alleles distributed in classes of base substitutions for cases and controls from the Goiânia population accidentally exposed to low doses of ionizing radiation.

A)


Nº Mendelian Deviation
Case
Control
2500
2000
1500
1000
Mendelian Deviation
Case
Control
Group

B)

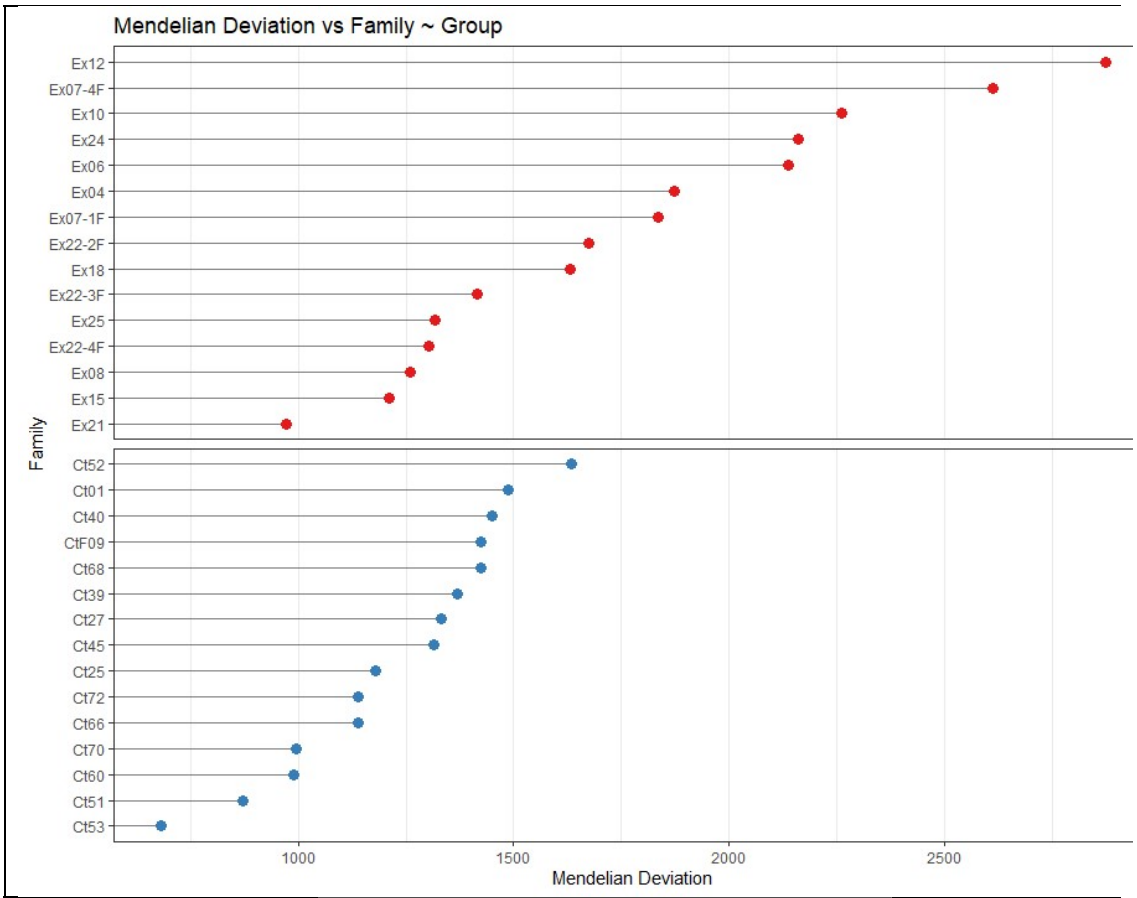


Fig 8. Representation of the variability of Mendelian deviations. (A) Mean numbers of Mendelian deviations in the offspring of the case and control groups. (B) Representation of the number of Mendelian deviations for each trio in the case and control groups.

SIPO: SCRIPTS FOR DETECTION AND ANALYSIS OF *DE NOVO* MUTATIONS FROM HIGH-DENSITY SNP GENOTYPING DATA

Hugo Pereira Leite Filho[1,2], Irene Plaza Pinto[3], Lorraynne Guimarães Oliveira[3,4], Emília Oliveira Alves Costa[3], Alex Silva da Cruz[3], Daniela de Melo e Silva[3,4], Claudio Carlos da Silva[1,2,3,5], Aparecido Divino da Cruz[1,3,4,5], Alexandre Rodrigues Caetano[6]

[1] Programa de Pós-Graduação em Biotecnologia e Biodiversidade, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[2] Universidade Estadual de Goiás, Goiás, Brazil
[3] Núcleo de Pesquisa Replicon, Mestrado em Genética, Escola de Ciências Agrária e Biológicas, Pontífice Universidade Católica de Goiás, Goiânia, Goiás, Brazil
[4] Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[5] Laboratório de Genética Molecular e Citogenética Humana, Secretaria de Saúde Pública do Estado de Goiás, Goiânia, Goiás, Brazil
[6] Embrapa Recursos Genéticos e Biotecnologia, Brasília, Distrito Federal, Brazil

Running title: **SIPO: DETECTION AND ANALYSIS OF DE NOVO MUTATIONS FROM SNP GENOTYPING DATA**

ABSTRACT

The authors developed SIPO, scripts inference parental origin, which are able to detect *de novo* mutations in trios of different sample groups, infer the origin of the parents in relation to the offspring, highlight the variant that has mutated and indicate the type of substitution in the nucleotides of the DNA strand from the SNP genotyping data analyzed with GeneChip® CytoScan HD™ 750K. The SIPO was applied to three distinct groups, only to validate the script. The sample contained 45 individuals (15 couples and 15 children). In the authors' study, they investigated the genotypes of each parent's SNP marker and compared them with the genotypes of their children. With the results obtained, it was possible to identify the mean rate of Mendelian deviation per group (μ_{MDA}±0.16%, μ_{MDB}±0.06%, μ_{MDC}±0.2%), in addition to deducing that the tool corroborates the issue of quality control of genotyping data, using as exclusion criteria the high numbers of Mendelian deviations observed in trios.

Keywords: Bioinformatics; SNPs; Genotyping; CytoScan 750K; *De novo* mutations.

INTRODUCTION

The development of computational tools to solve problems of biological origin has contributed in an innovative way to the advancement of science, especially in the area of applied genomics (HOGEWEG, 2011). With the sequencing of the human genome it has been possible to obtain an enormous amount of information on DNA structure and function, while millions of genetic variations observed between individuals have been unveiled. These findings opened new opportunities to identify the genetic basis of many traits, understand the causes of common diseases and reveal many evolutionary aspects of the human species

It is estimated that 90% of genetic variations among humans are SNPs - Single Nucleotide Polymorphisms (KOLKMAN et al., 2007; COLLINS et al., 1997). Also, variations in SNPs are important to determine genotypic and phenotypic relationships, within and between species and populations, and also to identify variants related to genetic diseases in humans and animals (Shah & Kusiak, 2004). For this reason, a great effort has been dedicated to the development of precise, fast, and cost-effective technologies for SNP analysis, resulting in many different approaches. Genotyping usually involves the generation of allele-specific PCR products among the SNPs of

interest, followed by detection with fluorescent probes to determine the genotypes (Kim and Misra, 2007). Commercial platforms provide high-density SNP arrays that can be used to genotype hundreds of thousands of markers in a single assay (Hernadez-Ferrer et al., 2015), such as the Thermo Fisher - CytoScan750 K HD™. Although Thermo Fisher Scientific provides an analysis suite such as the Chromosome Analysis Suite (ChAS™) with a wide range of functionalities to analyze the data obtained from their arrays, many custom-specific downstream analysis are not always contemplated and cannot be incorporated into the closed-platform software.

The objective of this study was to develop and validate a bioinformatics tool to answer a specific problem considering SNP data from a case-control study performed to compare numbers of *de novo* Mendelian deviations (DM) observed in trios of different sample groups. The develop tool allows to identify DM in trios, infer the parental origin of observed DM, identify the type of base substitution observed in the child, as well as to recognize mutated variants.

MATERIAL AND METHODS

Implementation

The developed pipeline termed SIPO (Scripts for Inference of Parental Origin) reads the ChAS® output file and performs an integrity check in the .CYCHP file. SNPQC (SNP Quality Control) is a measure of how well genotype alleles are resolved in the microarray data. Based on an empirical testing dataset, we have determined that array data with SNPQC < 15 (for CytoScan 750K and HD) is of poorer quality than is required to meet genotyping QC standards. Table 1 lists the variables generated by ChAS®. An initial filter is applied to each genotype call considering the Call Confidence (CC) parameter. Genotypes with CC $<5x10^{-2}$ are flagged and not considered in subsequent analyzes, as per manufacturer's guidelines. Remaining genotypes from each identified trio are subsequently compared, locus by locus, considering only situations where all three genotypes passed the initial filter. Only autosomal SNPs are considered in the analyzes to avoid previously reported issues associated with low coverage and with experimental noise observed on the Y and X chromosomes (Wise et al., 2013). Figure 1 shows a flowchart for all data processing steps performed by SIPO.

Applying solving recurrences consists of one or more base cases and one or more recursive cases. Each of these cases is an equation or inequality, with some function value *f (n)* on the left side. The base cases give explicit values for a typically finite, subset of the possible values of n. The recursive cases relate the function value *f (n)* to function value *f (k)* for one or more integers k < n. Typically, each recursive case applies to an infinite number of possible values of n, using equation 1: (Erickson, 1999)

$$f(x)= \begin{cases} 0 if n = 0 \\ 1 if n = 1 \\ f\left(\left[\frac{n}{2}\right]\right) + f\left(\left[\frac{n}{2}\right]\right) otherwise \end{cases} \qquad f(n) = \left\{ \begin{cases} 0 if n = 0 \\ 2xf\left(\frac{n}{2}\right) if n is even \wedge n > 0 \\ f(n-1)+1 \end{cases} \right\}$$

Outlined the recursion, in which each node of the problem represents the cost of n subproblems, being that the set of invocations determines a recursive function. This set of invocations is called the recursion tree (CORMEN et al., 2009). The solution to this theory was the algorithm known as mergesort, and the execution time of this algorithm is described by Equation 2 below (CORMEN et al., 2009; ERICKSON, 1999).

$$T(n) = aT\left(\frac{n}{b}\right) + f(n)$$

Where a≥1, b≥1 are constants; *T* is running time; n is number of problems; *f(n)* is root of the recursion tree.

Now *T(n)* is just the sum of all values stored in the recursion tree. For each i, the ith level of the tree contains a^i nodes, each with value $f(n=b^i)$. Thus, (KLEINBERG & TARDOS, 2006)

$$T(n) = \sum_{i=0}^{L} a^i f(n/b^i)$$

where L is the depth of the recursion tree. We easily see that $L = \log_b n$, because $\frac{n}{b^L} = 1$

Therefore, the input size is considered large because we have a number of possibilities between the variables (maternal and paternal genotypes are 2^{10} combinations for each one), which makes the order of growth in the execution time of SIPO relevant. In this way, we compare maternal genotypes (2^{10} combinations) vs. paternal genotypes (2^{10} combinations) and compare with offspring genotypes (2^9 combinations when the progenitors are homozygous and 2^7 when the parents are heterozygous).

All trio genotype calls observed to disagree from expected Mendelian inheritance rules are subsequently flagged. All flagged calls observed, representing potential Mendelian deviations (MD) are then analyzed for inferring parent-of-origin of the potencial germline mutation, and to classify the type of DNA base substitution observed (Table 2). Situations where it is not possible to identify the parental origin of an observed MD are marked with an "unsolved" tag. SIPO is also able to notify which variant has mutated. An output file is generated, including all original data in addition to all information contained in the original ChAS output file.

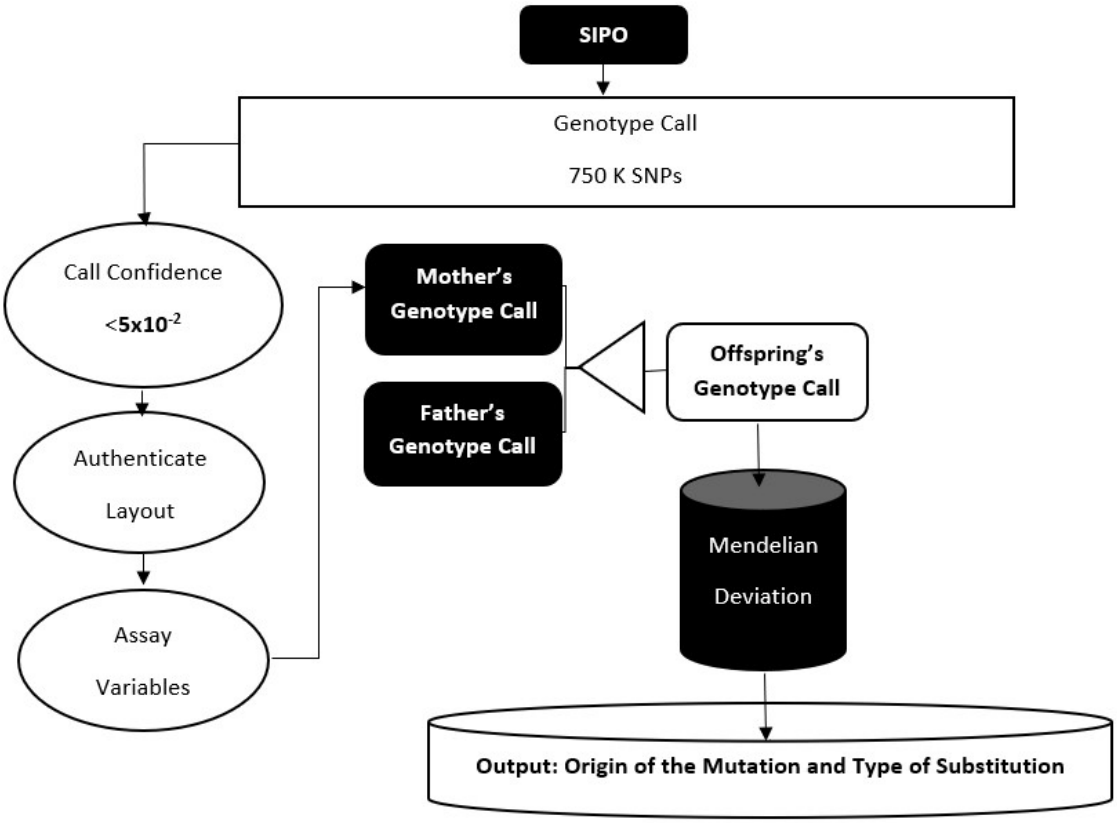


Figure 1. Representation of all processing steps performed by SIPO to identify potential *de novo* germline mutations observed as Mendelian deviations (MD) in analyzed trios, infer the parental origin of the MD, and classify the type of DNA base substitution.

Table 1. Variables taken the ChAS®-output for a sample trio and SNP locus.

Field	Description	Example
Marker	SNP marker identifier	S-3WRNV
Genotype Call	Trio genotype call in bi-allelic AB format (i.e.AA|AB|BB)	AA(Father)XAB(Mother) AB(Offspring)
Call Confidence	Confidence value for genotyping of each SNP (Father|Mother|Offspring)	1.754152E-14(Father) 4.44089E-16(Mother) 1.7919E-12(Offspring)
Signall A	Value representing the fluorescent intensity for the A channel for each genotype	499.7063(Father) 329.7947(Mother) 351.88184(Offspring)
Signall B	Value representing the fluorescent intensity for the B channel for each genotype	1601.1708(Father) 1675.4652(Mother) 1504.8402(Offspring)
Call Base	Trio genotype call in bi-allelic ATCG format	GG(Father)XGG(Mother) GG(Offspring)
dbSNP	SNP identifier record in NCBI database	rs2340582
Chromosome	Chromosome where marker is located	1
Position	SNP locus/cromosome base-pair position	882803

SIPO is implemented as Perl scripts freely available for download and use. Original scripts, codes, sample datasets, and obtained statistics are available in the github public repository at: https://github.com/hugofilho/sipo. Downstream statistical analysis of results generated with SIPO were performed using pipelines under R environment.

We applied the R package to further the interpretation of the results generated from the SIPO. Based on standard CRAN and Bioconductor classes, beyond packages DiagrammeR is a package for make flowchart, graph objects are rendered as SVG (Scalable Vector Graphics), the content is scalable and responsive to resizing of the plot window (Iannone, 2020); Companion to Applied Regression to applied Scatterplot Matrix (Fox & Weisberg, 2019); dplyr package is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges (Wickham, 2019); plotly was used by graph library makes interactive, publication-quality graphs (Sievert, 2020); ggalt (Rudis et al., 2017) and ggfortify (Tang et al., 2016) to contribute to the results of Principal Component Analysis; tidyverse (Wickham et al., 2019); rstatix (Kassambara, 2020) and ggpubr (Kassambara, 2020) makes to the packages work in harmony to clean, process, model, transform, building multiple t-test for different variables and visualize data.

Data Output

All observed MDs are subsequently processed to using R scripts to perform descriptive statistics, linear regression, and principal component analysis (spectral decomposition approach).

Table 2. Different scenarios considered in SIPO to infer the parent of origin of observed mutations.

Example	Father	Mother	Offspring	Parent-	Mutation	Mutation

				of-origin of observed MD	Type	Variant
1	AA	AA	AC	Unsolved	Transversion	Unsolved
2	AC	AA	CC	Mother	Transversion	A>C
3	AA	GG	CG	Father	Transversion	G>C

Software Validation

Samples from a case-control study were used for validating SIPO. The sample group (group A) was composed of five trios where one or both parents were exposed to ionizing radiation (0.2-0.5 Gy). The control group (group B) was composed of five trios where no individuals were exposed to ionizing radiation. All individuals in groups A and B had SNPQ ≥15. An additional group (group C) composed of five trios with no individuals exposed to ionizing radiation but with SNPQC <15. In total, samples from 45 individuals were used in the study. Biological paternity was confirmed for all trios.

All samples were analyzed with the high-density SNP genotyping microarray CytoScan HD™ 750K HD (Thermo Fisher, Thermo Fisher), following protocols provided by the manufacturer. The study was conducted at the Replicon Research Center of the School of Agricultural and Biological Sciences of the Pontifical Catholic University of Goiás in partnership with the Laboratory of Human Cytogenetics and Molecular Genetics/Public Health Laboratory Dr. Giovanni Cysneiros of the State Secretariat of Health of Goiás. SNP genotypes were called with the ChAS© (Thermo Fisher Scientific, Massachusetts, EUA) software version 3.1, using reference coordinates from the hg19 version of the human genome (University of California, Santa Clara's Genome Browser). For the analysis, the reference library CytoScanHD_Array.na33.r2.REF_MODEL and the analysis parameter file CytoScanHD_Array.na33.annot.db were used.

All analyzed data are in the public repository: https://github.com/hugofilho/sipo/tree/master/data/samples

RESULTS

The results were generated to demonstrate the execution of the SIPO, and we used the samples to validate the SIPO. The Table 3 contains the general and descriptive data about the analyzed samples.

Table 3 General data from groups A, B and C for the parental and F1 generations included in the study on the induction of germline mutation on offsprings.

Generation	Variables		A	B	C
Parental	N		5	5	5
	Age range (years)		19 to 56	20 to 40	21 to 42
	Mean age (years) at conception (±SD)	Father	32.8 (14.3)	32.2 (7.5)	31.4 (8.1)
		Mother	23.8 (4.5)	28.6 (7.5)	28 (4.5)
F1	N		5	5	5

	Mean MD (±SD)		2,408.80 (±182.05)	935.40 (±170.5)	3,557.40 (±457.2)
	Total MD		12,044	4,677	17,787

SD: Standard Deviation; M: Men; W: Women;

1824
1825
1826

In Table 4, the authors have the representation of the result generated by SIPO, indicating the MD, the parental origin (with inference or undetermined), type of substitution of the base.

$Name_{AFFY}$	$Name_{dbSNP}$	CHR	Position	Base Calls Father	Confidence Father	Base Calls Mother	Confidence Mother	Base Calls Offspring	Confidence Offspring	Mutation Origin	base substitutions	Variant
S-3JQLV	rs2485024	1	6199677	CC	1.044E-5	GG	8.339E-8	GG	0.002	Paternal	Transversion - Purine/Pyrimidine	C>G
S-4NUMZ	rs761108	1	6377743	GG	4.94E-9	AG	0.0	AA	0.0	Paternal	Transition - Purine/Purine	G>A
S-3YGAD	rs2898915	1	7063387	TT	4.844E-10	CC	0.004	TT	6.507E-4	Maternal	Transition - Pyrimidine/Pyrimidine	C>T
S-3QSCQ	rs4908646	1	7581993	GG	9.562E-7	AA	1.292E-12	AA	3.109E-11	Paternal	Transition - Purine/Purine	G>A

Table 4. Result of analyzing Dataset Ex24 with SIPO

In Table 5 contain the summary of the data used in this study, mutation burden was defined by the number of de novo base substitutions in an assayed SNP.

Table 5. Overrall data from groups used regarding the study based on the execution of SIPO

Family	Sex Parental Exposure	Absorbed Dose	father's age	Mother's age	Offspring's age	Parental Origin	Maternal Origin	Unsolved	Total MD*	Group
Ex06	Male	0.3	35	26	19	1,045	1,082	9	2,136	A

Ex07-4F	Female	0.2	56	20	17	1,509	1,085	18	2,612	A
Ex10	Female	0.2	21	24	2	1,187	1,054	21	2,262	A
Ex12	Female	0.3	31	30	3	1,361	1,486	28	2,875	A
Ex24	Male	0.5	21	19	12	1,010	1,102	47	2,159	A
Ct51	without exposure	0.0	35	34	2	384	479	8	871	B
Ct53	without exposure	0.0	35	27	8	315	361	6	682	B
Ct60	without exposure	0.0	40	38	0.9	501	482	6	989	B
Ct66	without exposure	0.0	31	20	26	543	594	2	1,139	B
Ct70	without exposure	0.0	20	24	8	448	545	3	996	B
CtS05	without exposure	0.0	22	21	9	1,562	1,600	15	3,177	C
CtS08	without exposure	0.0	26	26	25	1,974	1,666	85	3,725	C
CtS09	without exposure	0.0	30	31	10	1,814	2,317	130	4,261	C
CtS16	without exposure	0.0	42	31	5	1,506	1,603	48	3,157	C
CtS20	without exposure	0.0	37	31	6	1,700	1,755	12	3,467	C

*MD = Mendelian Deviation

In the present study, 49% (5,906), 16% (781), and 56% (9,902) of MD were identified, and their origins were indeterminate in groups A, B, and C, respectively (Figure 2).

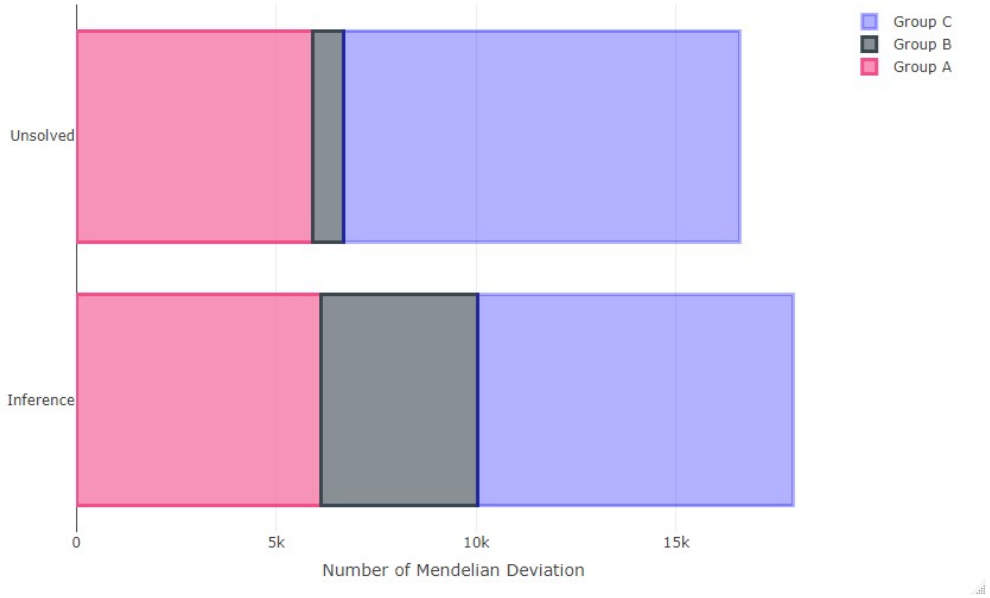


Figure 2. It indicates the proportion of inferences made by SIPO in groups A, B and C

We apply the obtained results by SIPO and it is possible to map the variations generated from the SNP variants mutated. Table 6 shows the values obtained between the groups.

Table 6. Summary of the descriptive data of the case and control groups for the six classes of base substitution in the genome of children conceived after parental exposure to low doses of ionizing radiation and their controls.

lGroup	Class	Minimum	Maximum	Mean	SD*	Total
A	C>A	60	111	84.71	23.73	422
	C>G	88	168	126.29	33.16	628
	C>T	285	573	427.71	115.35	2,136
	T>A	32	78	56	19.65	282
	T>C	322	549	438.71	105.60	2,200
	T>G	64	101	80.14	16.57	396
B	C>A	37	68	53.71	11.73	271
	C>G	39	102	69.86	24.85	348
	C>T	199	355	274.57	61.03	1,368
	T>A	34	40	36.57	2.51	182
	T>C	228	332	279.14	48.42	1,394
	T>G	47	72	57.71	10.93	285
C	C>A	75	159	107.00	34.29	535
	C>G	90	191	153.20	40.70	766
	C>T	329	770	520.00	188.52	2,600
	T>A	66	96	78.00	12.08	390
	T>C	352	776	585.60	161.40	2,928
	T>G	84	127	104.00	16.96	520

*SD=Standard Deviation

The figure 3 below shows the variants that have undergone mutations.

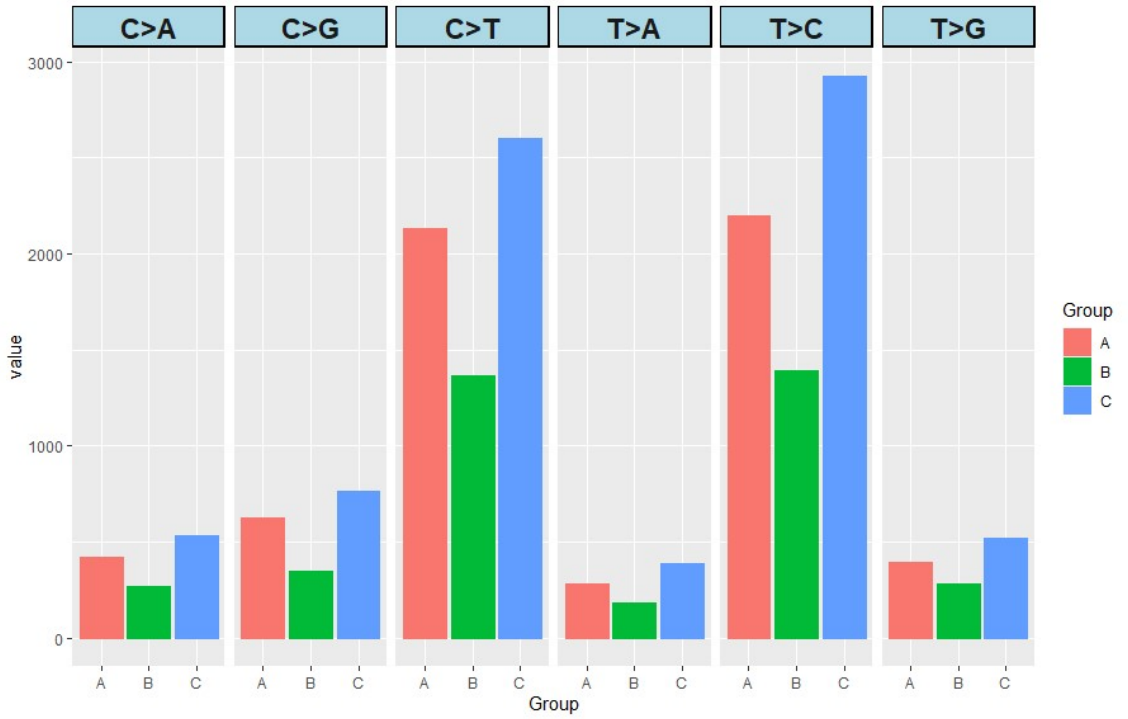
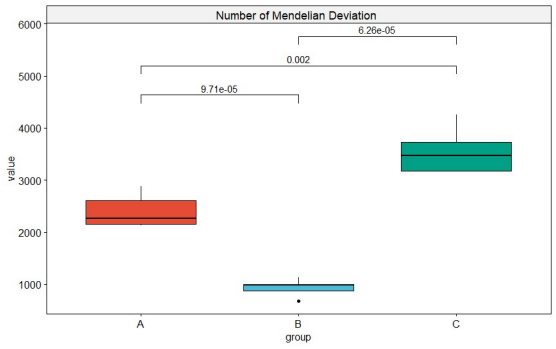


Figure 3. Phased de novo mutations based on parental and derived alleles distributed in classes of base substitutions for the groups A, B and C.

SIPO was used to identify MD differences in the three different analyzed groups. In Figure 4, the authors have the representation between the MD numbers for each group, the application of the T-test to demonstrate the significance ($p<5x10-2$, $\alpha=95\%$) in the difference between the mean MD numbers between the groups and the clustering of the groups.

A)

B)



C)

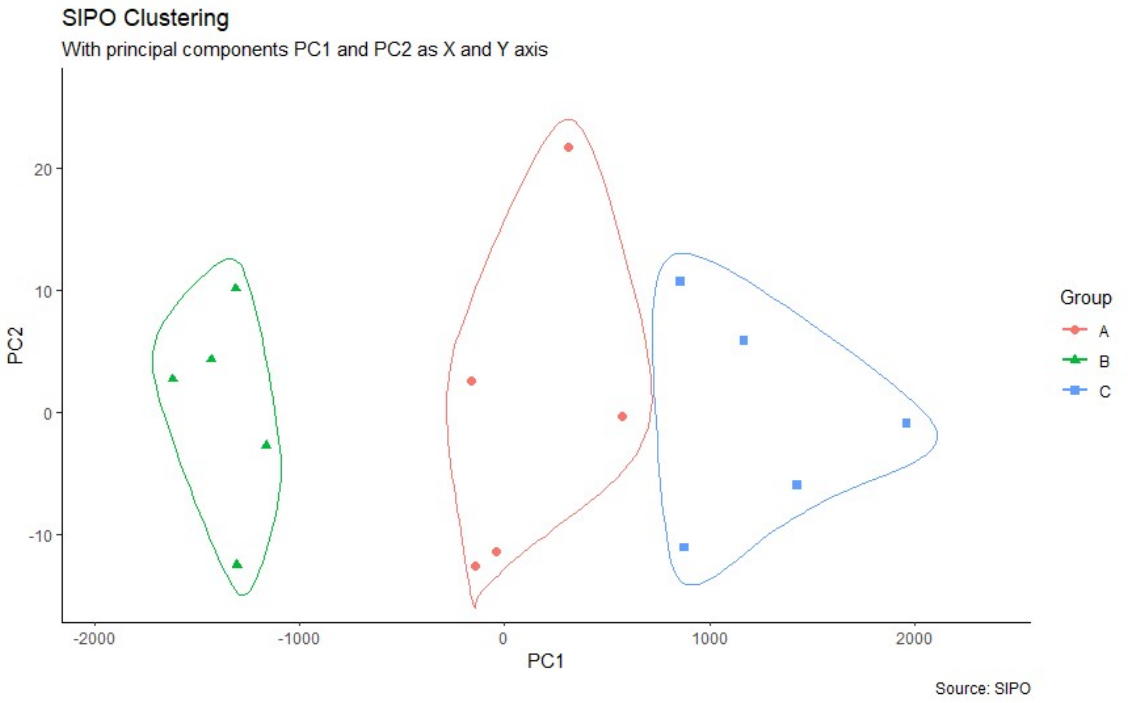


Figure 4. **a)** Representation of the MD numbers per group. **b)** Representation of the Mendelian deviation numbers using the T-test, with their respective p-values. **c)** Representation of the cluster of Mendelian deviation numbers separated by groups.

For age analysis, the authors applied the t-test statistical technique to compare a sample mean of maternal and paternal ages between groups. Figure 5 shows that the mean maternal and paternal ages were not significant, showing that the difference between the number of MD between the sample groups had no influence on the age factor.

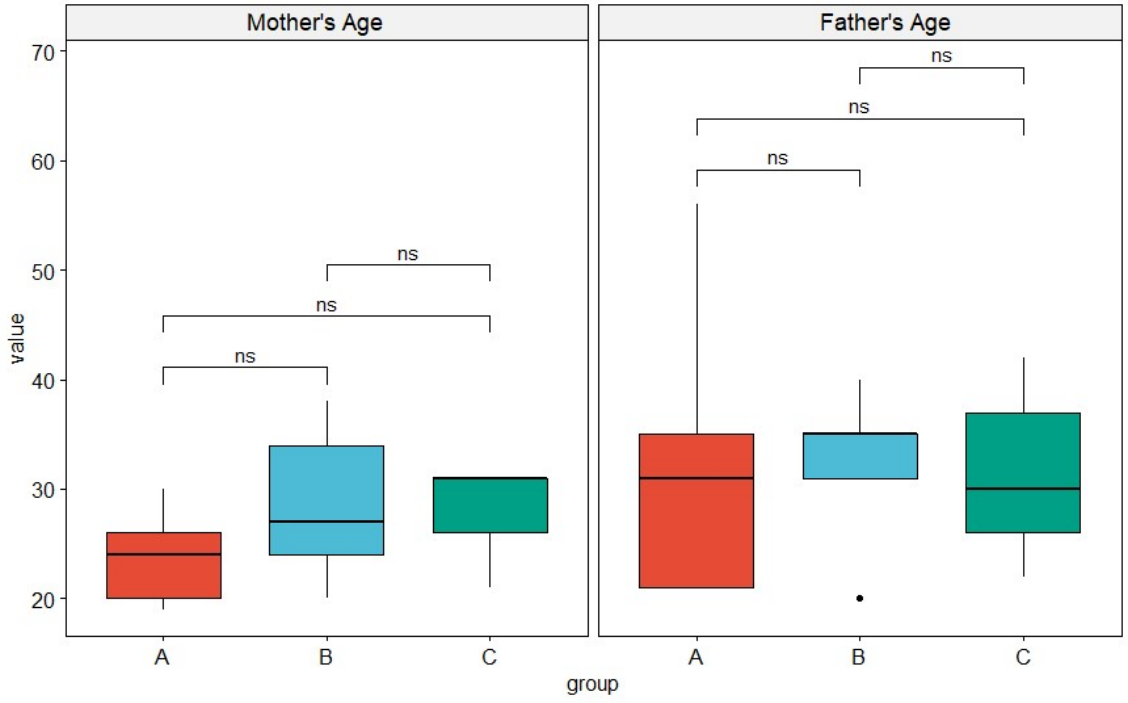


Figure 5. The graphs represent the comparison between the average ages of the parents in the different groups. *ns: non-significant.

DISCUSSION

In the present study, it was possible, from genomic data, to estimate retrospectively the number of Mendelian deviations, corresponding to germinative mutations of SNPs in different sample groups. The database treated in this study comes from another recent work to investigate losses and gains in CNVs induced by low doses of ionizing radiation (COSTA, et al., 2018). Thus, the hypothesis on quantifying the number of Mendelian deviations from the polymorphic markers in the database that generated the CNVs in the previous study was put forward. Therefore, the innovative character of the present study lies in the development of a pipeline to search and indicate the parental origin of MD in autosomal human SNPs obtained, mainly, from data from GeneChip® CytoScan microarray HD™. (Thermo Fisher Scientific, Massachusetts, EUA).

The solution implemented for the Mendelian *de novo* deviation findings was the creation of an algorithm (SIPO) to compare the genotyping data extracted from ChAS® to perform inferences regarding the origin of the mutation and the type of substitution of the base occurred in the offspring. Thus, we cover the analyzes generated from ChAS, becoming it an important tool in SNP studies.

The study analyzed the quality criteria of the data observing the recommendations required by the Manufacturer (MAPD $\leq$0.25; Waviness SD $\leq$0.12; SNPQC $\geq$15, Manual of the Enterprise Thermo Fisher Chromosome Analysis Suite 2.0 Software User). The authors performed a simulation with three distinct groups: in Group A, five trios were represented by a group that was exposed to low radiation exposure; in Group B, five trios without a history of radiation exposure were used; and in Group C, five trios without a history of radiation exposure were used, but had SNPQC <15.

By running the SIPO for mutation inference, the authors identified that the amount of Mendelian deviations was greater than expected (μ_{DMA}±0.16%), and the effect on ionizing exposure may have influenced this increase. On the other hand, group B generated a mean of approximately μ_{DMB}±0.06%, i.e., below the mean (μ_{DM} ~ 0.1%, Saunders *et al*., 2008). Group C presented μ_{DMC}±0.2%, representing the value above the average, considering the SNPQC<15, which is not recommended by the manufacturer, a characteristic that directly influenced the quality of genotyping with a high number of Mendelian deviations. The age factor was discarded as it did not present significant differences between the groups.

All the SNPs harboring C>T transitions in the data sets were not located in CpG islands and were all included in the analyses. Although all transitions and transversions were observed in our data set (Fig 3), C>T and T>C were overrepresented, for all groups, favoring the well-known hypothesis that human genome harbor a mutational bias toward A/T composition in the DNA stand (Lynch, 2010).

The MDs occurred randomly among the SNP markers in all groups (A, B and C), and all had a low frequency of calls. No markers with MD were identified that repeated themselves in a significant number, showing the reliability of the analyses. In these situations, SIPO presents itself as an effective algorithm to be used as an additional tool in defining the parental origin of polymorphic variants obtained by the microarray of SNPs.

In our results, we show that the age factor has no significant difference, thus, the age variable had no influence on the findings of Mendelian deviations *de novo* generated in the offspring.

CONCLUSION

The study allowed inferring the variation in Mendelian deviation findings that can be applied in different sample groups. Thus, with the results obtained, it was possible to

identify that the tool also corroborates in the quality control of genotyping data, using as exclusion criteria the high numbers of observed Mendelian deviations. In this way, it can be applied as a support resource to help improve the quality of the genotyping dataset, as it identifies Mendelian deviations, causing these markers to be excluded from the data mass, if needed.

Therefore, with the incorporation of deductions for decision making about the origin of MD in the offspring, SIPO was able to identify through inference or without inference the parental origin of the deviations observed, as well as to indicate the type of base substitution in this study. Finally, the authors were able to achieve their objectives and as a perspective, they will validate by other methodologies some *de novo* Mendelian deviations identified in the present study.

ACKNOWLEDGEMENTS

The authors wish to thank the voluntarily participation of the patients. This study was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPEG (Fundação de Amparo à Pesquisa do Estado de Goiás), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), D.M.S. and A.D.C. wish to thank for their scholarships. We also wish to express gratitute to CARA (Centro de Assistência ao Radioacidentado da SES-GO) for helping with contacting the group of exposed parents. Moreover, we thank Dr. Fernando Nodari for assisting with the issues from the software ChAS® (Thermo Fisher, USA). Lastly, we thank the volunteers who selfishly agreed to participate in the study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. A.R.C. is a CNPq research fellow.

BIBLIOGRAPHICAL REFERENCE

1.	Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. PLoS Comput Biol, 7(3), e1002021.
2.	Kolkman, J. M., Berry, S. T., Leon, A. J., Slabaugh, M. B., et al. (2007). Single nucleotide polymorphisms and linkage disequilibrium in sunflower. Genetics, 177(1), 457-468.
3.	Collins, F. S., Guyer, M. S., Chakravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science, 278(5343), 1580-1581.
4.	Kim, S., & Misra, A. (2007). SNP genotyping: technologies and biomedical applications. Annu. Rev. Biomed. Eng., 9, 289-320.
5.	Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature, 488(7412), 471-475.
6.	Acuna-Hidalgo, R., Veltman, J. A., Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome biology, 17(1), 1-19.
7.	Thermo Fisher "Data Sheet: The CytoScan® HD Cytogenetics Solution". Avaible [http://media.Thermo Fisher.com/support/technical/datasheets/cytoscan_hd_ datasheet.pdf]. Accessed 24 June 2020.a
8.	Thermo Fisher "Data Sheet: The CytoScan® 750K Cytogenetics Solution". Available at [http://media.Thermo Fisher.com/support/technical/datasheets/cytoscan750k_ datasheet.pdf]. Accessed 24 June 2020.b
9.	R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.
10.	Costa, E. O. A., Pinto, I. P., Gonçalves, M. W., da Silva, J. F., et al. (2018). Small de novo CNVs as biomarkers of parental exposure to low doses of ionizing radiation of caesium-137. Scientific reports, 8(1), 1-13.
11.	Manual do Enterprise Thermo Fisher® Chromosome Analysis Suite 2.0 TM Software User. Chromosome Analysis Suite 3.2 (ChAS 3.2). USER GUIDE. Publication Number 702943. Revision 11. ©2017 Thermo Fisher Scientific Inc.
12.	Saunders, I. W., Brohede, J., Hannan, G. N. (2007). Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. Genomics, 90(3), 291-296.
13.	Wise, A. L., Gyi, L., & Manolio, T. A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. The American Journal of Human Genetics, 92(5), 643-647.
14.	Richard Iannone (2020). DiagrammeR: Graph/Network Visualization. R package version 1.0.5. https://CRAN.R-project.org/package=DiagrammeR
15.	John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
16.	Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr
17.	C. Sievert. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC

	Florida, 2020.
18.	Bob Rudis, Ben Bolker and Jan Schulz (2017). ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'. R package version 0.4.0. https://CRAN.R-project.org/package=ggalt
19.	Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.
20.	Masaaki Horikoshi and Yuan Tang (2016). ggfortify: Data Visualization Tools for Statistical Analysis Results. https://CRAN.R-project.org/package=ggfortify
21.	Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
22.	Alboukadel Kassambara (2020). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.6.0. https://CRAN.R-project.org/package=rstatix
23.	Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
24.	Hernandez-Ferrer, C., Garcia, I. Q., Danielski, K., Carracedo, Á., Pérez-Jurado, L. A., & González, J. R. (2015). affy2sv: an R package to pre-process Affymetrix CytoScan HD and 750K arrays for SNP, CNV, inversion and mosaicism calling. BMC bioinformatics, 16(1), 167.
25.	Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., ... & 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.
26.	Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences. 2010; 107(3): 961-968.
27.	KLEINBERG, Jon; TARDOS, Eva. Algorithm design. Pearson Education India, 2006.