Journal of Biotechnology and Biodiversity

journal homepage: *https://sistemas.uft.edu.br/periodicos/index.php/JBB/index*

Conflito de interpretação e desequilíbrio de ligação em variantes encontradas nos genes BRCA1 e BRCA2

Caio Agostini Calheiros Grosso[a]*, Tiago César Gouvêa Moreira[a], Luciana de Andrade Agostinho[b]

[a] Fundação Cristiano Varella, Brasil
[b] Centro Universitário UNIFAMINAS, Brasil
* Autor correspondente (caioagostiny@gmail.com)

INFO

Keyworks
bioinformatics
BRCA1 and BRCA2
linkage disequilibrium

ABSTRACT

Interpretation conflict and linkage disequilibrium in BRCA1 and BRCA2 genetic variants.
The aim of this study was to analyze BRCA1 and BRCA2 genetic variants in order to compare *in silico* predictions by different tools. In addition, we investigated the linkage disequilibrium (LD) in the genetic variants observed in patients with cancer. ClinVar, SIFT, PolyPhen2, VEP, PROVEAN and Fathmm-MKL were performed for *in silico* analysis. The investigation of the linkage disequilibrium was performed with Linkage Disequilibrium Calculator software. We observed 60 different variants and the most common was T>C. The programs SIFT, PolyPhen2 and PROVEAN showed similarities in the degree of pathogenicity observed in variants. VEP, Fathmm-MKL and ClinVar predicted the majority of variants analyzed. Sixteen genetic variants, manually selected, were analyzed as LD and 33 of them were confirmed when analyzed in pairs, then, were joined in 5 groups. Genetic variants observed in our sample are usually observed in others populations as in South America, in South Asia and East Asia, Africa, and Europe. Correct association of phenotype/genotype and epidemiological information can provide important epidemiological information, such as prognostic and treatment aspects, seeking a better quality of life, understanding of diseases and genetic evolution factors associated.

Palavras-chaves
bioinformática
BRCA1 e BRCA2
desequilíbrio de ligação

RESUMO

Este estudo teve como objetivo analisar variantes encontradas nos genes *BRCA1* e *BRCA2* para comparar as diferentes predições *in silico* de ferramentas distintas, além de investigar variantes candidatas ao desequilíbrio de ligação (DL). Para as análises *in silico* foram utilizados os programas ClinVar, SIFT, PolyPhen2, VEP, PROVEAN e Fathmm-MKL. A investigação do DL foi feita pelo programa *Linkage Disequilibrium Calculator*. Observou-se 60 variantes distintas, predominando as trocas nucleotídicas de T>C. Os programas SIFT, PolyPhen2 e PROVEAN apresentaram semelhanças quanto ao grau de patogenicidade determinado em cada variante. O VEP, o Fathmm-MKL e o ClinVar apresentaram resultado de predição para maior parte das variantes analisadas. A partir das 16 variantes com suspeita de DL, selecionadas manualmente, foram confirmados 33 DL quando analisadas aos pares. A combinação das variantes em DL resultou na categorização de 5 grupos. As variantes genéticas observadas em nossa amostra são encontradas com maior frequência na América, Sul da Ásia e Leste Asiático, na África e na Europa. A associação do fenótipo do paciente com seu genótipo e as informações epidemiológicas das variantes encontradas no câncer, assim como informações sobre o prognóstico e tratamento associados, podem proporcionar melhor qualidade de vida, entendimento das doenças e de fatores evolutivos associados.

Received 09 September 2020; Received in revised from 20 September 2020; Accepted 30 April 2021

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

INTRODUÇÃO

O câncer é considerado um dos grandes problemas da saúde pública em todo o mundo. No Brasil, as estimativas referentes ao número de novos casos de câncer foi de 625 mil em 2020. Sendo o câncer de mama, o tipo mais incidente e, o câncer de ovário, o sétimo mais frequente em mulheres (INCA, 2019).

Nas últimas décadas, têm-se investigado os mecanismos moleculares pelos quais mutações genéticas adquiridas ou herdadas conseguem transmitir as suas células filhas, vantagens para escapar dos controles normais de crescimento e diferenciação celular. Alterações nas proteínas envolvidas no ciclo celular, nos oncogenes e em genes supressores tumorais podem subsidiar o desenvolvimento de neoplasias (Alldredge e Randall, 2019).

Os genes *Breast Cancer 1* (*BRCA1*) e *Breast cancer 2* (*BRCA2*) são classificados como supressores tumorais, cujas proteínas codificadas participam no reparo, na replicação e na transcrição do DNA (Sousa *et al.*, 2019).

Ambos os genes conferem susceptibilidade para o câncer de mama, de ovário, de próstata, de pâncreas, de estômago e de colorretal, e a alta proporção do câncer de mama e ovário familial está associada a mutações encontradas nestes genes (Sousa et al., 2019; Scott et al., 2019).

Nos últimos anos, houveram grandes avanços no cenário da Biologia Molecular, como o desenvolvimento e implantação de tecnologias de sequenciamento de nova geração (NGS). Embora esta ferramenta seja de fácil acesso em laboratórios de genética, interpretar os dados obtidos tem sido um desafio no campo da bioinformática (Scott et al., 2019 ; Madroñero *et al.*, 2019)

A informática tornou-se essencial para a área da biologia, permitindo reconhecimentos e análises de sequências genéticas, além de predições para configuração tridimensional de proteínas e identificação de inibidores de enzimas. A bioinformática ainda auxilia na organização e na relação com a informação biológica, no agrupamento de proteínas homólogas, na análise de experimentos de expressão gênica e no estabelecimento das árvores filogenéticas (Linhares, 2014; Borges et al., 2020).

As variantes, do tipo SNP, mutações pontuais, são utilizadas em estudos de associação e mapeamento genético, em ensaios diagnósticos para confirmação de paternidade, identificação individual e detecção de doenças genéticas (Caetano, 2009).

As variantes encontradas podem proporcionar informações que auxiliam estudos populacionais e de ancestralidade. Sendo o desequilíbrio de ligação (DL), uma análise complementar entre as variantes quando se busca este tipo de investigação (Jorde, 2000).

Além das interpretações clínicas, as variantes podem ser analisadas em banco de dados e plataformas *in silico*. Há uma diversidade em relação a interpretação e análise das mesmas, inclusive entre os diferentes bancos de dados e ferramentas utilizadas (Linhares, 2014).

Algumas ferramentas de bioinformática que podem realizar esta análise *in silico* são: o *Sorting Intolerant From Tolerant* (SIFT), o *Polymorphism Phenotyping* (PolyPhen2), o *Variant Effect Predictor* (VEP) (Linhares, 2014), o Fathmm-MKL (Ferçaino et al., 2017) e o *Protein Variation Effect Analyzer* (PROVEAN) (Montenegro *et al.*, 2021).

Devido à grande quantidade de dados gerados e interpretados de formas distintas, existe um desafio na determinação do grau de patogenicidade destas variantes, uma vez que estes dados são interpretados por curadores diferentes e por utilizarem técnicas de predições diferentes em relação às ferramentas *in silico*. Isso faz com que a análise destas variantes se torne laboriosa e de alta complexidade (Sarmento, 2013).

Este estudo teve como objetivo realizar a análise *in silico* das variantes *SNP* observadas nos genes *BRCA1* e *BRCA2* de pacientes com câncer, utilizando diferentes programas *in sílico*. Investigou-se também as relações de DL entre as variantes, as quais foram separadas em grupos, associação com o fenótipo dos pacientes e comparação de sua frequência em diferentes populações do mundo.

MATERIAL E MÉTODOS

Aspectos éticos da pesquisa

A pesquisa foi aprovada pelo CEP-FAMINAS (CAAE:62262416.3.0000.5105).

Casuística

As variantes SNPs analisadas são provenientes de pacientes com câncer de mama, ovário, colorretal e próstata do Hospital do Câncer de Muriaé-MG atendidos no primeiro semestre de 2018. As variantes foram investigadas pela ferramenta Sequenciamento de Nova Geração (NGS) pelo sistema *Ion Torrent* (*Thermo Fisher Scientific*) no laboratório de Biologia Molecular deste hospital. Os genes *BRCA1* e *BRCA2* foram analisadospela técnica de sequenciamento massivo do tipo NGS.

Critérios de inclusão e exclusão

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

As variantes (SNP) que foram incluídas para a análise *in silico* foram encontradas em pacientes com câncer de mama, de ovário, de colorretal e de próstata do Hospital do Câncer de Muriaé – MG. As variantes obtidas apresentaram cobertura de leitura *(reads)* > 25, em homozigotos, com Variação de Frequência Alélica (VAF) entre 90 a 100%, e *reads* > 50, em heterozigotos, com VAF de 40 a 60%. Assim, as variantes que apresentaram *reads* e VAF fora das condições mencionadas, foram excluídas (conforme validação prévia deste protocolo por *Sanger* para exclusão das variantes falso positivas).

As variantes do tipo *InDel* não foram incluídas, uma vez que todas foram falso positivas quando reanalisadas por sequenciamento do tipo *Sanger*.

Análises das variantes por bioinformática

Realizou-se a análise *in silico* das variantes genéticas, encontradas nos genes *BRCA1* e *BRCA2* pela plataforma *National Center for Biotechnology Information* (NCBI), a qual permitiu visualizar a sequência correta de bases nitrogenadas dos genes. As alterações genéticas cadastradas e disponíveis no ClinVar e NCBI também foram observadas (versão genômica GRCh:37.p13).

Na plataforma, *Exome Aggregation Consortium* (ExAC) obteve-se a frequência das variantes em diferentes populações do mundo.

As variantes tiveram o grau de patogenicidade preditas pelos *softwares*: SIFT, PolyPhen2 e VEP, elucidados como ferramentas de bioinformática para predições por Linhares (2014), além do Fathmm (Sousa et al., 2019) e PROVEAN (Montenegro et al., 2021).

Em relação aos graus de patogenicidade, as variantes podem ser interpretadas de acordo com os seguintes critérios: benignas, *Variant of Uncertain Significance* (VUS), patogênicas e *De novo* (Richards et al., 2015).

Investigação do desequilíbrio de ligação das variantes SNP

As variantes que aparecem em maior frequência e em conjunto entre os pacientes, por tipo de câncer, foram verificadas para desequilíbrio de ligação pela plataforma do *Ensembl*. Esta plataforma calcula a existência de DL e proporciona dois coeficientes para esta análise: o r^2 (coeficiente de correlação entre os dois *loci*) e o *D'* (coeficiente de desequilíbrio de ligação).

O DL foi analisado de acordo com as populações com maiores frequências alélicas de cada variante (Figura 1).

Uma das populações utilizada foi a CEU (Residentes de Utah com Ancestrais da Europa do Norte e Ocidental) devido à grande frequência de imigrantes europeus recebidos no Brasil (Valverde *et al.*, 1958).

Outras populações em que as variantes tiveram sua frequência investigada foram as: da América, da África, do Sul da Ásia e do Leste Asiático, devido à alta frequência destas variantes nestas populações conforme os bancos de dados.

Figura 1 - Fluxograma de análise do DL das variantes encontradas.

A presença de desequilíbrio de ligação entre os *loci* foi considerada somente quando o $D' \neq 0$, caso contrário, o equilíbrio de ligação foi considerado (Borges *et al.*, 2020).

Análise estatística

Os dados foram tabulados em Excel (versão 2013) e foi realizada análise estatística descritiva e comparação de médias pelo IBM SPSS (versão 17)
.

RESULTADOS E DISCUSSÃO

Amostra

Neste estudo, 36 pacientes com diferentes tipos de câncer foram investigados, e suas variantes foram analisadas *in silico* e para .

Dentre as 118 variantes analisadas, 58 foram eliminadas por serem repetidas (Figura 2).

Dentre as outras 60 variantes, apenas 16 foram testadas para desequilíbrio de ligação por aparecerem com alta frequência e em conjunto em pacientes diferentes. As 16 variantes geraram, em combinação, 33 análises de desequilíbrio de ligação quando combinadas em pares.

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

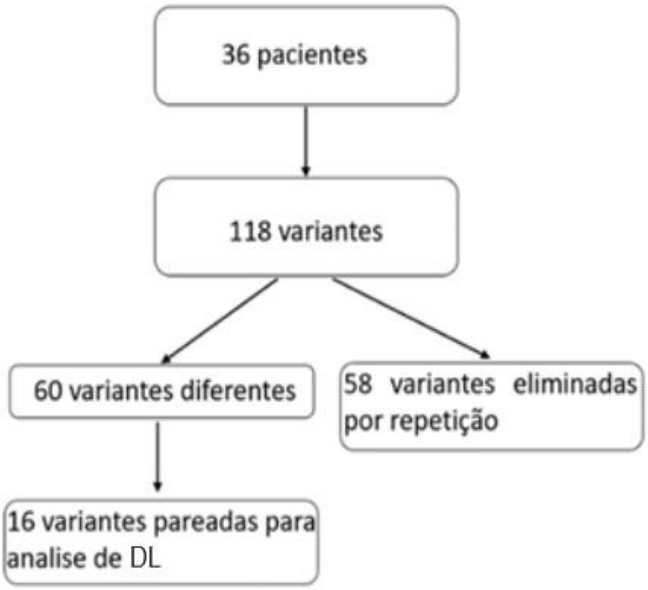

Figura 2 - Seleção e análise das variantes obtidas dos 36 pacientes.

As 118 variantes encontradas nos pacientes também foram associadas com cada tipo de câncer (Figura 3).

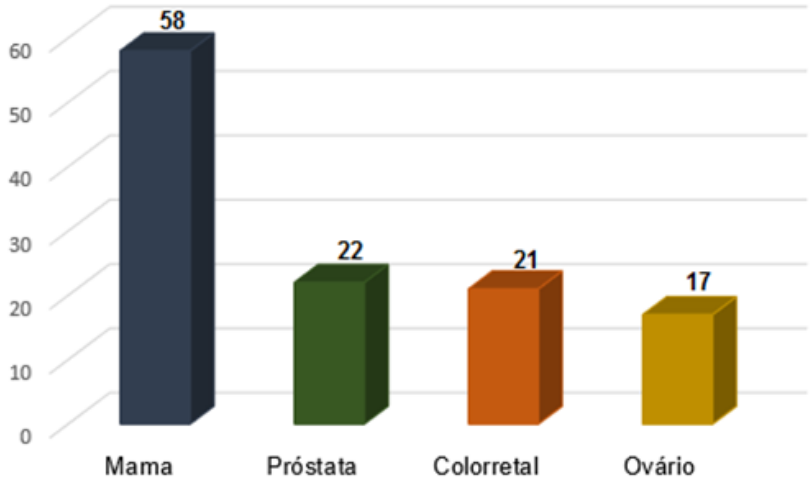

Figura 3 - Relação da quantidade de variantes por tipo de câncer.

Dentre as variantes relatadas, 58,3% se encontram em exons e, as outras, em introns.

Nos exons de ambos os genes, foram observadas 33,3% de variantes do tipo não sinônimas e, 21,6%, variantes sinônimas. Em 3,3% das variantes, foram observadas alterações em regiões de *splicing*, 1,6% em regiões 5'UTR (não traduzidas) e 1,6% em códons de parada.

As alterações mais frequentes no DNA envolvem bases nitrogenadas de mesma característica estrutural, isto é, troca entre duas purinas (G/A ou A/G) ou entre duas pirimidinas (T/C ou C/T) (Kwok e Gu, 2000).

As principais trocas nucleotídicas encontradas foram a de timina por citosina (T>C), adenina por guanina (A>G), citosina por timina (C>T) e guanina por adenina (G>A), em ordem decrescente (Figura 4).

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

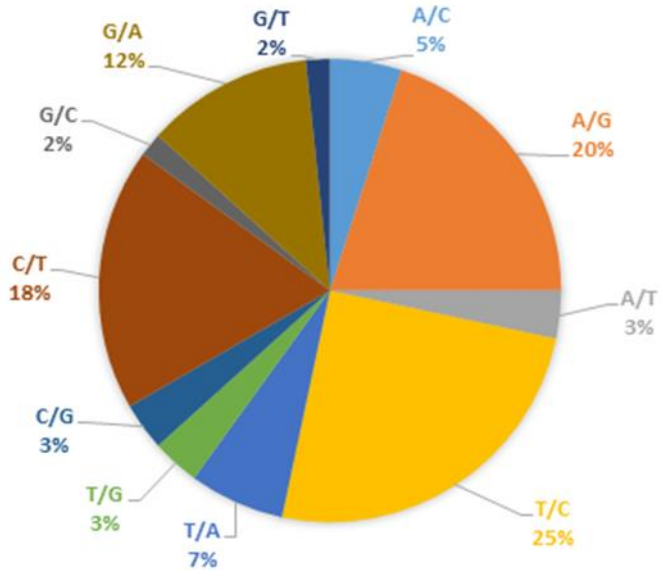

Figura 4 - Frequência relativa das trocas nucleotídicas encontradas nos genes *BRCA1* e *BRCA2*.

As trocas nucleotídicas de T>C foram encontradas em 25% das variantes genéticas, sendo que 53,3% destes SNPs ocorreram em regiões de exons, 40% em introns e 6,7% em sítios de *splicing*. As alterações em regiões de *splicing* ocasionam modificações no tamanho dos exons que podem alterar a estabilidade do RNA mensageiro (RNAm) ou da própria proteína sintetizada a partir deste RNAm (Viana, 2017).

Dentre 18% das variantes analisadas com trocas C>T, 27,3% ocorreram nos exons e o restante em introns. Todos os SNPs com troca A>G estavam presentes em exons, entretanto, 8,3% dessas variantes encontravam-se no códon de parada.

Os polimorfismos únicos que causam supressão e códons de parada de tradução, geram proteínas de diferentes tamanhos (Viana, 2017). Nas alterações nucleotídicas de G>A, 71,4% aconteceram em exons e, 28,6% em introns.

A substituição dos aminoácidos devido aos SNPs podem ocasionar modificações na sequência proteica, assim alterações estruturais e funcionais poderão influenciar num maior potencial de reflexo biológico que possibilita o desenvolvimento de diversas doenças e/ou síndromes (Kwok, 2000; Viana, 2017).

Interpretação clínica das variantes - Variantes interpretadas pelo banco de dados ClinVar

O ClinVar é um banco de dados hospedado pelo *National Center for Biotechnology Information (NCBI)* que permite a busca de informações do genótipo e fenótipo das variantes encontradas nos pacientes. Os dados são emparelhados com os do *Exome Aggregation Consortium* ou do *Genome Aggregation Database* (gnomAD), para a interpretação das variantes (Zhang et al., 2017).

O ClinVar é alimentado pela publicação de artigos científicos realizados principalmente por ensaios *in vitro* e *in vivo*. Neste estudo, este banco de dados viabilizou a classificação de 78,3% das variantes como benignas, 18,3% como VUS e 3,4% como novas ou sem classificação conhecida. O ClinVar não classificou nenhuma das variantes deste estudo como patogênica.

Além disso, as variantes encontradas no ClinVar, podem ter suas interpretações complementadas por meio da análise *in silico* de outras ferramentas, para um melhor entendimento do comportamento de estruturas moleculares dentro de um sistema de componentes interativo (Michels *et al.*, 2019).

Comparação entre as diferentes ferramentas preditoras

As variantes obtidas pelo sequenciamento NGS foram analisadas por diferentes plataformas *in silico*, citadas previamente neste estudo. Observou-se divergências classificatórias em relação ao grau de patogenicidade das mesmas (Tabela 1) e isso pode ocorrer devido às diversas metodologias usadas por cada *software* na análise de SNPs.

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

Tabela 1 - Predições *in silico* das variantes encontradas em pacientes com câncer de mama, de ovário, de cólon e de próstata realizada por ferramentas diferentes.

Predições	SIFT (%)	PolyPhen2 (%)	PROVEAN (%)	VEP (%)	Fathmm-MKL (%)
Benigna	33,4	38,4	48,3	73,3	80,0
Patogênica	18,3	8,3	3,4	6,7	15,0
VUS	-	-	-	6,6	-
Sem resultado	48,3	53,3	48,3	13,4	5

As classificações oriundas do ClinVar assemelham-se com as da ferramenta Fathmm-MKL em relação à predição das variantes como benigna e, em apenas 1,7%, estas predições foram distintas. Foi observada maior divergência quando foi comparada a interpretação obtida pelo ClinVar e as obtidas pelos softwares de análises *in silico* no SIFT, PolyPhen2 e PROVEAN.

Por meio de ferramentas preditoras *in sílico*, foi possível observar que mais de 50% das variantes, dos diferentes tipos de câncer, foram classificadas, com exceção do PolyPhen2, que classificou apenas 46,7% das variantes.

As variantes com grau de patogenicidade incerto, denominadas como *Variant of Uncertain Significance* (VUS), foram obtidas apenas pelo VEP e ClinVar. Todos os programas tiveram problemas em classificar pelo menos 3 variantes, as quais foram classificadas como VUS ou não obtiveram qualquer informação de predição (sem resultado).

Em relação à classificação *in silico* das variantes como patogênicas, os programas SIFT e Fathmm-MKL demonstraram frequências semelhantes. Assim como, o VEP e o PolyPhen2, que reportaram percentuais próximos na predição de variantes patogênicas. A predição *in silico* de variantes patogênicas variou de 3,4% a 18,3% dentre as ferramentas utilizadas.

O SIFT, PolyPhen2 e PROVEAN possuem classificações semelhantes, com diferença de 5% na determinação do grau de patogenicidade. Os resultados do PROVEAN, para previsões de substituição de um único aminoácido, são semelhantes aos resultados obtidos pelo SIFT e PolyPhen2, conforme observado no estudo de Choi e Chan (2015), em que 52,8% dos polimorfismos comuns foram previstos corretamente pelas ferramentas. Entretanto, ainda existem variantes em que os três programas geram previsões diferentes (Choi, 2012).

Segundo Choi (2012), o PROVEAN apresenta vantagem sobre as demais ferramentas SIFT e PolyPhen2, pois além do *software* de Busca por Alinhamento Local Básico (BLAST), ele também utiliza uma abordagem de pontuação de alinhamento delta para gerar uma matriz. Esta por sua vez, surge a partir de cálculos e da captura implícita de informações sobre a frequência de substituição e das propriedades químicas de 20 resíduos de aminoácidos. Além disso, a pontuação pode ser influenciada pela região que circunda o local da variante genética analisada.

O SIFT e o PolyPhen2 apresentam semelhanças quanto a quantidade de variantes classificadas, isto deve-se a semelhança existente entre ambas as ferramentas computacionais, por utilizarem o processo de alinhamento múltiplo de sequência por meio do software BLAST (Dakal, 2017).

Entretanto, ambas as ferramentas apresentam diferenças nas frequências das variantes preditas como benignas e patogênicas, ocasionando resultados falso positivos e falso negativos.

As variações de classificação entre o SIFT e o PolyPhen2 ocorrem devido ao método pelo qual cada programa utiliza para analisar o resíduo de aminoácido ou a sequência proteica de interesse.

O SIFT usa a ferramenta *Position Specific Iterated* BLAST (PSIBLAST) para busca de sequências proteicas similares e o alinhamento múltiplo de todas estas sequências para criar uma matriz de valores de posição específica. Assim, a partir da matriz, calculam-se todas as probabilidades para as substituições de resíduos de aminoácidos possíveis para cada posição do alinhamento, gerando, ao final, um score (Islam et al., 2018; Vaser et al., 2016).

Já o PolyPhen2, igualmente ao SIFT, usa o programa BLAST, e a partir dele, calcula uma matriz de *scores*. A diferença principal deve-se ao fato deste *software* analisar a estrutura da proteína, algo não avaliado pelo SIFT (Vaser et al., 2016).

Deste modo, a posição do resíduo de aminoácido variante é mapeada utilizando as informações das bases de dados *SWALL* e *SwissProt* com a avaliação da possibilidade de afetar o núcleo hidrofóbico da proteína, de acessibilidade a solvente, de interações eletrostáticas e de interações com ligantes (Adzhubei *et al.*, 2013).

Conforme Choi (2012), o VEP e Fathmm-MKL possuem maior capacidade de predição, quando comparados as demais ferramentas. Isto ocorre devido a uma deficiência do SIFT e PolyPhen2 em predizerem variantes quando encontram poucas ou nenhuma sequência homóloga para alinhamento.

Neste estudo, o VEP predisse 86,6% das alterações genéticas. A elevada capacidade de predição

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

das variantes pelo VEP ocorre devido à análise do impacto da variação em um transcrito ou proteína em banco de dados como o GENCODE (Harrow et al. 2012) e a Sequência de Referência (RefSeq) no NCBI (Pruitt et al., 2014).

O VEP dispõe de uma quantidade maior de bancos de dados para gerar informações quanto ao grau de patogenicidade das variantes genéticas, por exemplo, em variantes de RNAs não codificadoras, em regiões reguladoras ou em locais para ligação de fatores de transcricionais e locais de conservação. Além disso, as frequências alélicas encontradas neste *software* são provenientes do banco 1000 genomas, do NHLBI exoma e do ExAC (McLaren et al., 2016).

Os identificadores do PubMed e do ClinVar, possibilitam exclusão de variantes comuns como patogênicas quando associam as mesmas aos achados da literatura em relação às informações de significância clínica. Mesmo assim, existem variantes que quando analisadas *in silico* são deletérias e em evidências científicas não estão associadas como patogênicas em estudos in vivo (McLaren et al., 2016).

É importante ressaltar que as variantes preditas como patogênicas por ferramentas *in silico* devem possuir estudos *in vitro* ou *in vivo* que confirmem o efeito patogênico sobre o gene ou o sobre o produto gênico (Shihab et al., 2015).

O VEP fornece indicação do efeito da alteração de aminoácidos usando propriedades biofísicas de proteínas, com geração de pontuações e previsões pelo SIFT, PolyPhen2, Condel, Fathmm e Mutation Taster (McLaren et al., 2016). Deste modo, pelo fato de o VEP utilizar informações geradas no SIFT e no PolyPhen2 para avaliação da patogenicidade, o mesmo deveria retornar predições iguais as observadas no SIFT e PolyPhen2.

Entretanto, o VEP não apresentou predições semelhantes com as ferramentas SIFT e PolyPhen2, o que resultou na classificação de 6,6% das variantes analisadas pelo VEP como VUS, resultando nos conflitos de patogenicidade entre estes programas.

O Fathmm-MKL proporcionou a predição de 95% das variantes, a alta capacidade para realizar estas análises deve-se ao fato desta ferramenta não utilizar o BLAST, e sim, utilizar os múltiplos aprendizados de Kernel (MKL). Os diferentes tipos de dados sejam estes contínuos, discretos, de sequência ou de gráfico, são codificados em matriz de Kernel (Shihab et al., 2015).

Para prever se uma variante SNP é funcional ou não, utiliza-se um classificador baseado no aprendizado múltiplo do Kernel, que resulta em várias matrizes numéricas grandes que representam os diferentes tipos de bancos de dados. Cada um deles determina diferentes escalas de mensuração de acordo com a natureza da região genética analisada (Shihab et al., 2015).

Com o MKL, cada tipo de dado é codificado em um Kernel base correspondente $K\ell$ (o qual $\ell = 1$, e p se houver p grupos de recursos), a qual derivará a matriz composta do kernel $K = \sum p\ell = 1\lambda\ell K\ell$ em que os $\lambda\ell$ são pesos do núcleo que são ≥ 0. Desta forma, a matriz pode ser usada como um classificador baseado em Kernel juntamente com uma máquina de vetor de suporte (SVM) (Shihab et al., 2015).

A maioria das plataformas usa o BLAST para procurar, em bancos de dados, sequências homólogas de proteína para obter uma previsão computacional baseada na conservação evolutiva das sequências ou das anotações estruturais proteicas. Sendo assim, os programas como o SIFT e PolyPhen2 que utilizam esta metodologia, acabam gerando padrões de resultados semelhantes pré-definidos, por usarem a mesma metodologia baseada em uma matriz de posição específica que atribui pontuações diferentes a substituição do aminoácido, dependendo da posição onde ela ocorre (Reeb et al., 2020).

Entretanto, as classificações por ferramentas como o Fathmm usam do modelo de Markov (HMMs) associado aos métodos de Kernel (Fathmm-MKL). O método probabilístico utilizado pelo Fathmm-MKL é mais rigoroso por captura de informações de posições específicas dentro de um alinhamento de múltiplas sequências e, por detectar relações distantes entre as sequências homólogas (Shihab *et al.,* 2013).

Desequilíbrio de ligação

A análise das variantes dos genes *BRCA1* e *BRCA2* em 36 pacientes, permitiu a seleção manual de variantes a serem estudas para análise de eventuais desequilíbrios de ligação, por serem observadas em conjunto em pacientes diferentes e com alta frequência dentro da amostra deste estudo.

A análise da existência do DL resultou em 33 associações de regiões genéticas quando 16 variantes foram selecionadas e analisadas aos pares pelo software *Linkage Disequilibrium Calculator* (Figura 5).

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

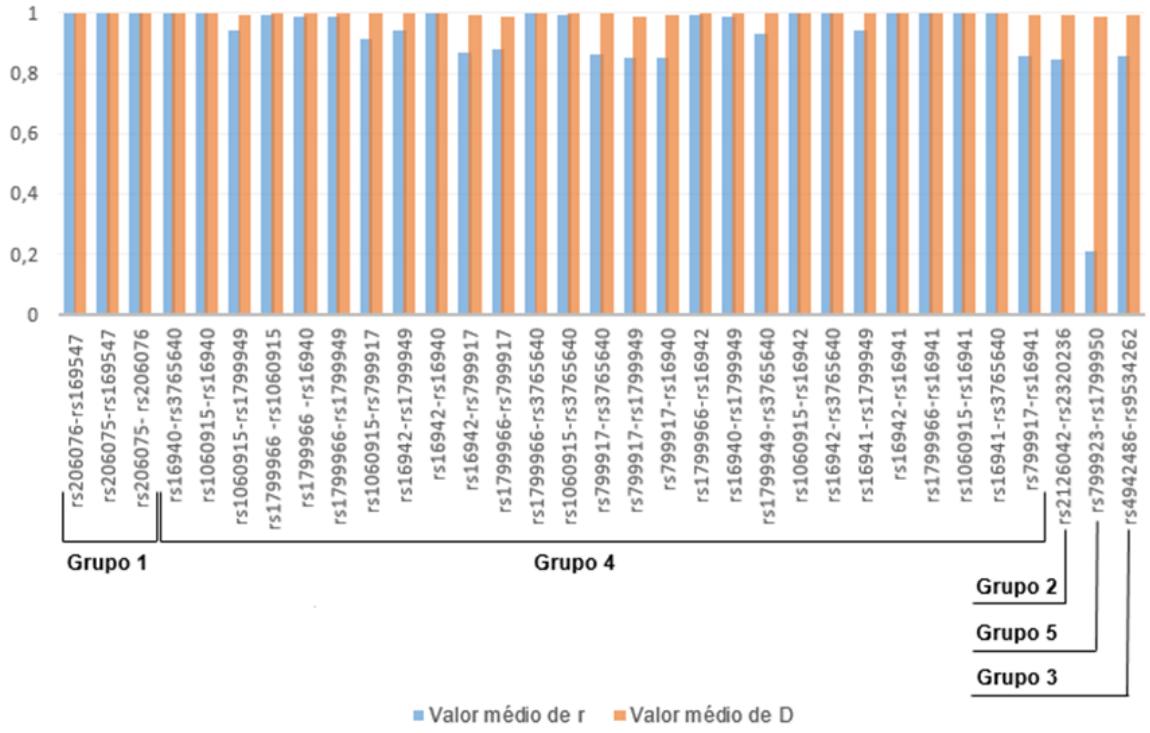

Figura 5 - Desequilíbrio de ligações observadas entre os pares de variantes nos genes *BRCA1* e *BRCA2*.

No grupo de pacientes com câncer de mama e próstata foi possível visualizar os 33 pares de variantes em desequilíbrio de ligação, no câncer de colorretal 31 das DL estavam presentes e, no de ovário, apenas 4.

O desequilíbrio de ligação é uma associação não aleatória entre alelos de diferentes loci em uma determinada população (Sadeghi et al., 2018). Sendo um fenômeno genético que interfere na dinâmica genética das populações, a qual possui efeito observável pela segregação não independente dos alelos dos diferentes loci, ocasionando correlação entre eles, durante a formação dos haplótipos e no momento da transmissão dos genes (Borges et al., 2020).

A partir de marcadores bialélicos, informações estatísticas podem ser usadas para expressar o DL , por meio do valor absoluto da diferença existente entre as frequências gaméticas observadas e esperadas sob equilíbrio de ligação gênica (D), a proporção de D em relação ao valor máximo na população (D') e o quadrado da correlação entre os valores de alelos de dois loci (r2) (Sadeghi et al., 2018; Flint-Garcia et al., 2003).

A existência de desequilíbrio de ligação entre os loci é considerada somente quando o D' ≠ 0, caso contrário, tem-se ligação de equilíbrio. Desta forma, os estudos de desequilíbrio de ligação e dos desvios presentes no Equilíbrio de Hardy-Weinberg, pode-se demonstrar o quanto e quais forças evolutivas atuam sobre determinada população (Borges et al., 2020).

Em relação às variantes com desequilíbrio de ligação, 33,3% apresentaram $r^2 = 1$ e D' = 1, demonstrando desequilíbrio de ligação absoluto por apresentarem baixa ou nenhuma diferença entre as frequências alélicas observadas.

Sendo assim, este desequilíbrio de ligação ocorre quando os polimorfismos correlacionados ocorrem em um período semelhante e nenhuma recombinação dentre as variantes (Flint-Garcia et al., 2003). Observações imediatas a respeito do desequilíbrio de ligação podem ser realizadas quando se tem um D' e r^2 altos, que indica a possibilidade de apenas dois haplótipos estarem presentes e com pequena variabilidade alélica em uma população. Os altos valores de D' e baixo r^2 exemplificam a possibilidade de estarem presentes em mais de um par de haplótipos, com pares de alelos com frequências distintas, refletindo a mesma história recombinacional, mas, diferentes trajetórias mutacionais (Duggal et al., 2019).

Neste estudo, 60,6% das regiões em DL analisadas tinham $r^2 < 1$, que segundo Flint-Garcia e colaboradores (Flint-Garcia et al., 2003), há ocorrência de recombinações nas linhagens alélicas.

Após as análises dos pares de regiões genéticas, cada par foi manualmente combinado a outros pares em DL, para que mais variantes fossem inseridas em um único grupo. Todas as variantes em DL encontradas formaram ao final 5 grupos (Figura 6).

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

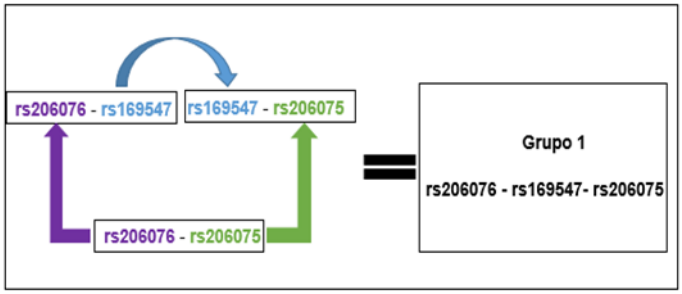

Figura 6 - Fluxograma com as associações feitas manualmente entre os pares de variantes em DL até a formação de cada grupo.

No gene *BRCA2* foram observados três grupos de variantes e, em *BRCA1*, dois grupos (Tabela 2), em que todas elas estavam presentes nos pacientes de forma combinada.

Tabela 2 - Grupos de variantes em LD nos genes *BRCA1* e *BRCA*.

Genes	Grupos	Variantes
BRCA2	Grupo 1	rs206075, rs169547 e rs206076
	Grupo 2	rs2126042 e rs2320236
	Grupo 3	rs4942486 e rs9534262
BRCA1	Grupo 4	rs16940, rs3765640, rs1060915, rs1799949, rs1799966, rs1899917, rs16942 e rs16941
	Grupo 5	rs799923 e rs1799950

O desequilíbrio de ligação presente entre as variantes genéticas que compõem o grupo 4 foram também relatados e classificados por Cox e colaboradores (Cox et al., 2011), como haplogrupo B.

Algumas variantes que compõem os grupos 1, 2 e 3 já foram relatadas em outros estudos, porém, não foram associadas em haplótipos, e sim, ao tipo de câncer (García-Perdomo et al., 2019; Freedman et al., 2004).

As duas variantes do grupo 5 também foram relatadas no estudo de Douglas e colaboradores (2011). Conforme estes autores, elas compõem um bloco em desequilíbrio de ligação (rs1799966, rs3737559, rs799923 e rs1799950) em pacientes com câncer de próstata. Em nosso estudo, estavam presentes em pacientes com câncer de Mama e Próstata.

O desequilíbrio de ligação, encontrado entre as variantes dos diferentes tipos de câncer, pode estar relacionado a diversos fatores evolutivos, tais como: o tempo decorrido da fundação da população ou os eventos mutacionais, o tamanho e o crescimento populacional, além de outros eventos genéticos (Borges et al., 2020).

As variantes analisadas em DL não estão associadas ao câncer neste estudo, pois não foram encontradas variantes patogênicas.

O aumento da frequência do DL das regiões pode ocorrer devido ao isolamento genético entre linhagens, ao pequeno tamanho populacional, à subdivisão populacional, à baixa taxa de recombinação, às misturas populacionais e à seleção natural e artificial (Sadeghi et al., 2018).

Um dos principais eventos demográficos que influencia o DL é a deriva genética, que é um processo evolutivo que altera a frequência dos alelos ao longo das gerações devido ao acaso, efeito fundador ou efeito gargalo. Assim, quanto menor a população, maior o indício da deriva genética, com aumento de DL (Kliman et al., 2008)

Com a expansão de uma população, os efeitos da deriva genética e das mutações tornam-se menores, assim como a distância observada entre dois *loci* em desequilíbrio (extensão) entre todos os pares de *loci* de uma geração para outra. Em contrapartida, rápidas expansões acabam por limitar o desequilíbrio de ligação mediante a diversidade populacional e recombinações, além da amenização da deriva genética (Jorde, 2000).

Em uma população pequena, pode-se encontrar variação significativa no desequilíbrio de ligação entre dois *loci*, decorrente das taxas de mutações, uma vez que elas podem alterar frequências alélicas e haplotípicas (Jorde, 2000).

Os eventos de miscigenação dependem de proporções e níveis de diferenças nas frequências alélicas das populações ancestrais, do tempo decorrido desde a miscigenação e das contribuições da população parental para quantificar e identificar a extensão do DL (Kliman et al., 2008).

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

A seleção natural pode proporcionar o surgimento de desequilíbrio de ligação, ao modo que, alelos novos para atingirem altas frequências populacionais, necessitam de muitas gerações de existência. Contudo, quando há seleção positiva têm-se o aumento da frequência do alelo selecionado em poucas gerações, tendo redução de recombinação e do número de haplótipos existentes (Andrade, 2013).

A queda da frequência de regiões em DL é ocasionada pela ocorrência de elevadas taxas de recombinação em uma linhagem, além de mutações e redução repentina do tamanho da população (Sadeghi et al., 2018).

Associações das variantes em grupos e o tipo de câncer

Apesar das variantes encontradas não estarem associadas ao fenótipo dos pacientes analisados (tipo de câncer), os grupos de variantes foram relacionados com eles (Tabela 3).

Tabela 3 - Relação da quantidade de pacientes por tipo de câncer categorizados pelos grupos de variantes.

Grupo de variantes	Genes	Tipos de câncer	Quantidade de pacientes
1	*BRCA2*	Mama	26
		Ovário	1
		Próstata	3
		Colorretal	6
2	*BRCA2*	Mama	7
		Colorretal	2
		Próstata	1
3	*BRCA2*	Mama	16
		Ovário	1
		Próstata	1
4	*BRCA1*	Mama	16
		Próstata	2
		Colorretal	1
5	*BRCA1*	Mama	3
		Próstata	1

O grupo com câncer de mama foi o que teve mais indivíduos investigados (n=26). As associações existentes entre as variantes do grupo 1 foi a única que esteve presente em todos os 36 pacientes e nos quatro tipos de câncer.

De acordo com a tabela 3, 50% dos pacientes com câncer de ovário (n=2) contém o grupo 3 ou 1 de variantes. Dentre os pacientes com câncer de colorretal (n=6), 16,6% estão contidos no grupo 1 e 4 e, 33,3%, de todos os pacientes, estão contidos nos grupos 1 e 2.

Durante a análise *in silico* do DL, observou-se que 92,3% pacientes com câncer de mama apresentaram pelo menos dois grupos de variantes associadas.

Dos pacientes com câncer de mama, 41,7% continham dois grupos de variantes, 45,8% apresentaram três grupos e 12,5%, quatro grupos com associações entre os pares com DL. Dos pacientes com câncer de próstata, 33,3% possuem os três grupos (1, 2 e 3) no gene *BRCA2* e um grupo no *BRCA1*, além de 66,4% deles estarem nos grupos 1 e 4.

Observou-se que, mesmo tendo tipos de câncer diferentes, os pacientes compartilham os mesmos grupos de variantes. Isso pode ocorrer pela semelhança dos pacientes em relação ao seu perfil genético associado a origem ancestral, pois são pacientes residentes na Zona da Mata Mineira. Este perfil genético pode ser reflexo do processo de colonização do Brasil e dos fluxos migratórios para esta região de Minas Gerais (Valverde *et al.*, 1958).

Associação das variantes e sua frequência em diferentes populações

Uma vez que as variantes encontradas em DL podem ser utilizadas em estudos de origem ancestral para análise populacional (Andrade, 2013), os cinco grupos obtidos tiveram suas variantes associadas a diferentes populações do mundo.

É importante ressaltar que a semelhança do perfil genético das variantes investigadas neste estudo com outras populações, não associa o fator causal do câncer a sua origem ancestral, uma vez que não foi encontrada mutação patogênica pelo ClinVar nos pacientes.

As variantes encontradas foram comparadas às populações da África, da América, da Europa, do

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

Leste da Ásia e do Sul da Ásia. Estas regiões foram selecionadas como contexto da análise de DL devido ao fato das variantes encontradas, neste estudo, estarem em alta frequência nestas populações (Tabela 4).

Tabela 4 - Comparação das variantes encontradas nas populações da África, da América, da Europa, do Leste da Ásia e do Sul da Ásia e análise de DL .

Grupos	1ª População mais frequente		2ª População mais frequente	3ª População mais frequente	DL (1ª População mais frequente)	DL (2ª População mais frequente)	DL (3ª População mais frequente)
Grupo 1	rs206076	América	Sul da Ásia	Leste Asiático	Sim	Não	Não
	rs206075	América	Sul da Ásia	Leste Asiático	Sim	Não	Não
	rs169547	América	Sul da Ásia	Leste Asiático	Sim	Não	Não
Grupo 2	rs2126042	Africana	América	-	Sim	Sim	-
	rs2320236	Africana	América	-	Sim	Sim	-
Grupo 3	rs4942486	América	-	-	Não	-	-
	rs9534262	Africana	-	-	Sim	-	-
Grupo 4	rs16940	Sul da Ásia	Africana	-	Sim	Não	-
	rs3765640	Sul da Ásia	Africana	-	Sim	Não	-
	rs1060915	Sul da Ásia	Africana	-	Sim	Não	-
	rs1799949	Sul da Ásia	Africana	-	Sim	Não	-
	rs1799966	Sul da Ásia	Africana	-	Sim	Não	-
	rs799917	Africana	Sul da Ásia	-	Não	Sim	-
	rs16942	Sul da Ásia	Africana	-	Sim	Não	-
	rs16941	Sul da Ásia	Africana	-	Sim	Não	-
Grupo 5	rs799923	Europa	-	-	Sim	-	-
	rs1799959	Europa	-	-	Sim	-	-

As variantes contidas no grupo 1, apresentaram frequências alélicas idênticas para as três populações (América, Sul da Ásia e Leste Asiático) com 100%. Somente na América, foi possível encontrar desequilíbrio de ligação entre os pares de variantes que representam o grupo 1.

As altas taxas de recombinação dentro de uma população, faz com que o desequilíbrio de ligação seja reduzido rapidamente a cada geração, formando outros haplótipos. Com o passar do tempo, o equilíbrio pode ser atingido depois de algumas gerações. Assim, sugere-se que as populações do Sul da Ásia e do Leste Asiático tenham passado por eventos recombinantes que propiciaram a redução do desequilíbrio de ligação entre as variantes do grupo 1.

As duas variantes que fazem parte do grupo 2 possuem maior frequência na população Africana: rs2126042 (27,5%) e rs2320236 (23,1%) e na Americana: rs2126042 (22%) e rs2320236 (21%). Em ambas as populações, foi possível evidenciar desequilíbrio de ligação entre as variantes. É possível que as taxas de recombinação nestas populações não sejam elevadas (D' = 0).

O grupo 3 tem variantes que apresentam, em separado, maiores frequências alélicas em populações diferentes, sendo assim, o rs4942486 tem maior frequência na América (54%), sem DL e o rs9534262, tem maior frequência na África (58%) com DL quando associado ao rs4942486.

A miscigenação pode ocasionar o surgimento de novos haplótipos com queda do DL, já no momento da fundação de uma população pode gerar um alto DL (Andrade, 2013; Scott et al., 2019).

Foi observado DL entre as variantes do grupo 4 somente na população do Sul da Ásia eAfricana, e de acordo com a alta frequência, é a população mais semelhante. Todas as variantes que possuem maior frequência no Sul da Ásia apresentam frequência alélica de 50%.

O rs799917, diferente dos outros SNPs, exibe maior frequência alélica (88,5%) na África, e não no Sul da Ásia (53%). Entretanto, o rs799917 não está em desequilíbrio de ligação com as demais variantes que compõem o grupo 4 na África, apenas no Sul da Ásia, apesar de possuir maior frequência

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

na África.

As variantes do grupo 5 apresentaram maior frequência alélica na população da Europa quando comparadas às outras populações, rs799923 (22,7%) e rs1799950 (6%), sugerindo perfil genético semelhante a esta população. Também foi possível evidenciar o DL na população Europeia.
Na população americana, também foi observado DL entre o par de variantes do grupo 5, mesmo sendo a segunda população mais frequente apenas para a variante rs1799950.

A colonização da zona da Mata Mineira ocorreu principalmente pela Europa com forte movimento migratório para o Brasil (Valverde et al., 1958). Apesar da Europa não ser o local com alta frequência das variantes encontradas nos grupos 2, 3 e 4, todas elas estavam em desequilíbrio de ligação nesta população.

Apenas no grupo 5, a Europa foi a população com maior frequência das variantes em DL. As variantes em DL encontradas no câncer de mama e de próstata possuem altas frequências na Ásia e na África.

Os pacientes com câncer colorretal e de ovário, tiveram variantes que são encontradas nas populações da América, Ásia e África, mediante a análise de grupos por câncer e do estudo de ancestralidade das variantes que compõem os mesmos.

A população brasileira é formada por uma mistura de quase cinco séculos entre colonizadores europeus, escravos africanos e ameríndios, atribuindo uma formação tri-híbrida. Assim, os processos migratórios entre as regiões brasileiras durante o período colonial promoveram a dispersão e a miscigenação destas três principais populações em diferentes proporções ao longo do país (Tarazona et al., 2015).

As características da estrutura genética da população brasileira são evidenciadas pela colonização europeia, sendo marcado pela alta frequência de casamentos de homens de ancestralidade europeia e mulheres com ancestralidade africana ou nativa. Por meio de estudos de patrilinearidade, os brasileiros possuem origem ancestral europeia com maior frequência e, seus marcadores de matrilinearidade tem maior frequência na origem ancestral africana e nativo americana (Tarazona et al., 2015).

A migração é um fator evolutivo que promove a introdução de novos alelos da população migrante para a população receptora, assim a população receptora se tornará mais semelhante à população de origem dos migrantes (Godinho, 2008).

As associações de desequilíbrio podem surgir por meio do fluxo gênico, assim ocorre o aparecimento de associações alélicas entre dois *loci* que não estão ligados. Estas associações decaem com o tempo e dependem da distância genética entre os grupos parentais, contudo o desequilíbrio de ligação diminui de acordo com as taxas de recombinação recorrente em várias gerações de cruzamento (Godinho, 2008; Sigrist, 2012).

O fluxo gênico é um dos eventos evolutivos que justificam a miscigenação presente em grande parte dos indivíduos da população brasileira, que por meio de combinações genéticas principalmente entre africanos, europeus e nativos americanos, geram diferentes haplótipos (Tarazona et al., 2015).

Desta forma, compreender o processo descrito e discernir a população de origem do desequilíbrio de ligação ajudará no mapeamento de genes relacionados a doenças e à evolução das populações (Sigrist, 2012).

CONCLUSÕES

Dentre as 60 variantes analisadas, a troca de bases pirimidínicas T>C foi a mais frequente neste estudo, com predominância em exons.

As análises realizadas pelas ferramentas *in silico* resultaram em algumas predições diferentes, mas em sua maior parte, os programas SIFT, PolyPhen2 e PROVEAN apresentaram semelhanças em seus resultados.

Se comparado ao VEP, Fathmm-MKL e ClinVar, as ferramentas SIFT, PolyPhen2 e PROVEAN tiveram maior percentual de variantes sem classificação (sem resultado), com mais de 40% das variantes. Estas três ferramentas (SIFT, PolyPhen2 e PROVEAN) utilizam metodologias diferentes em relação ao VEP e o Fathmm-MKL.

O ClinVar foi a única ferramenta que não classificou nenhuma das variantes como patogênica, ademais, o ClinVar possui interpretações semelhantes com as do VEP e do Fathmm-MKL nas predições das variantes genéticas benignas.

Algumas variantes foram preditas como patogênicas pelas ferramentas *in silico* SIFT, PolyPhen2, PROVEAN, VEP e Fathmm-MKL, e para estas mesmas variantes, o ClinVar reportou como variante com significado incerto (VUS). Para estas variantes são necessários ensaios *in vitro* e *in vivo* para que mais informações sejam obtidas.

A ferramenta *in silico* que foi mais adequada em nosso estudo foi o Fathmm-MKL, por possuir um método analítico que considera diferentes tipos de dados e por retornar *scores* baseados na matriz de Kernel.

Ao interpretar variantes genéticas, os profissionais da saúde devem ter cautela, visto que, cada ferramenta apresenta uma referência ou metodologia diferente para se determinar o grau de patogenicidade. Para isso, seria importante categorizar cada programa pelo tipo de referência

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

e/ou método utilizado, para se analisar cada variante e determinar o grau de patogenicidade utilizando resultados comparáveis em um mesmo contexto de análise. Além de ser importante a associação com a literatura.

Em relação às variantes encontradas, por tipo de câncer, em conjunto e com alta frequência, todas tiveram o DL confirmado.

Quando os 5 grupos de variantes foram determinados, observou-se maior semelhança com as populações da América, África e Ásia, em ordem decrescente. Pacientes com câncer de mama e de próstata tiveram predominância das variantes que são semelhantes com as da população da Ásia e da África.

Já os pacientes com câncer colorretal e de ovário, tiveram predominância com as variantes que são semelhantes às populações da América, Ásia e África. A semelhança do perfil genético das variantes investigadas neste estudo com outras populações, não confirma a associação do fator causal do câncer a sua origem ancestral. Mais estudos e pacientes com variantes patogênicas são necessários para esta análise.

A correta associação do fenótipo do paciente com seu genótipo pode proporcionar informações epidemiológicas importantes, assim como informações sobre o prognóstico e tratamento, em busca de melhor qualidade de vida e de um maior entendimento das doenças e de fatores evolutivos dentro de uma população.

AGRADECIMENTOS

Agradecemos ao Hospital do Câncer de Muriaé por terem fornecido as variantes genéticas e ao departamento de genética do INCA (Instituto Nacional de Câncer) pela validação por Sanger das variantes.

REFERÊNCIAS BIBLIOGRÁFICAS

Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet, v.7, n.20, jan. 2013.

Alldredge J, Randall L. Germline and Somatic Tumor Testing in Gynecologic Cancer Care. Obstet Gynecol Clin North Am, v.46, n.1, p.37-53, 2019.

Andrade CCF. Estrutura genética e desequilíbrio de ligação em africanos, ameríndios e remanescentes de quilombos brasileiros estimados por novos STRs-X. 2013. Tese (Doutorado em genética) – Faculdade de Meicina de Ribeirão Preto da Universidade de São Paulo, Ribeirão Preto: SP.

Borges MG, Rocha CS, Carvalho BS, Lopes-Cendes I. Methodological differences can affect sequencing depth with a possible impact on the accuracy of genetic diagnosis. Genetics and Molecular Biology, v.43, n.2, 2020.

Caetano AR. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. Rev Bras Zootec, v.38, p.64-71, 2009.

Choi Y, Chan AP. Provean web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics, v.31, n.16, p.2745-2747, 2015.

Choi Y. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. ACM, New York, p.414-417, 2012.

Cox DG. *et al*. Common variants of the BRCA1 wild-type allele modify the risk of breast cancer in BRCA1 mutation carrier. Hum Mol Genet, v.20, n.23, p. 4732-47, 2011.

Dakal TC, Kala D, Dhiman G, Yadav V, Krokhotin A, Dokholyan NV. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in *IL8* gene. Scientific reports, v.7, n.6525, 2017.

Douglas JA, Levin AM, Zuhlke KA, RAY AM, Johnson GR, Lange EM, Wood2 DP, Cooney KA. Common variation in the BRCA1 gene and prostate cancer risk. Cancer Epidemiol Biomarkers, v.16, n.7, p.1510–1516, 2011.

Duggal P, Ladd-Acosta C, Ray D, Beaty TH. The evolving field of genetic epidemiology: From familial aggregation to genomic sequencing. American Journal of Epidemiology, v.188, n.12, p.2069–2077, 2019.

Ferlaino M, Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR; Campbell C. Open access an integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. BMC Bioinformatics, v. 18, n.442, 2017.

Figueiredo JC, Brooks JD, Conti DV, Poynter J. N, Teraoka SN, Malone KE, Bernstein L, Lee WD, Duggan DJ, Siniard A, Concannon P, Capanu M, Lynch CF, Olsen JH, Haile RW, Bernstein JL. Risk of contralateral breast cancer associated with common variants in BRCA1 and BRCA2: Potential modifying effect of BRCA1/BRCA2 mutation carrier status. Breast Cancer Res Treat, v.127, n.3, p.819–829, 2011.

Flint-Garcia SA, Thornsberry J M, Buckler ES. Structure of linkage disequilibrium in plants. Annual Review Plant of Biology, v.54, p.357-374, 2003.

Freedman ML, Penney KL, Stram DO, Marchand LL, Hirschhorn JN, Kolonel LN, Altshuler D, Henderson BE, Haiman CA. Common variation in BRCA2 and breast cancer risk: a haplotype-based analysis in the multiethnic cohort. Human Molecular Genetics, v.13, n.20, p.2431–2441, 2004.

García-Perdomo HA, Saldarriaga MAB, Sánchez A. Frequency of variants in DNA-repair genes in a southwest colombian population. Urol Colomb, v.28, n.03, p.226-233, 2019.

Godinho NMO. O impacto das migrações na constituição genética de populações latino-americanas. 2008. Tese (Doutorado em Ciências Biológicas), Universidade de Brasília, Brasília.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans, M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE Project. Genome Res, v.22, n.9, p.1760–74, 2012.

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso

Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2020: incidência de câncer no Brasil / Instituto Nacional de Câncer José Alencar Gomes da Silva. – Rio de Janeiro: INCA, 2019.

Islam SMA, Kearney CM, Bake EJ. Assigning biological function using hidden signatures in cystinestabilized peptide sequences. Scientific reports**,** v.8, n.9049, 2018.

Jorde LB. Linkage disequilibrium and the search for complex disease genes. Genome Res, v.10, n.10, p.1435-1444, 2000.

Kliman R, Sheehy B, Schultz J. Genetic drift and effective population size. Nature Education, v.1, n.3, p.3, 2008.

Kwok P, Gu Z. Single nucleotide polymorphism libraries: why and how are we building them? Molecular medicine today, v.5, n.12, p.538-43, 2000.

Linhares ND. Aplicação clínica do sequenciamento e análise bioinformática de exoma. 2014. Tese (Doutorado em Genética), Universidade Federal de Minas Gerais, UFMG.

Madroñero LJ, Corredor-Rozo ZL, Escobar-Pérez J, Velandia-Romero ML. Next generation sequencing and proteomics in plant virology: how is Colombia doing?. Acta biol. Colomb. v.24, n.3, p.423-438, 2019.

Mclaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. Genome Biology, v.1, n.17, p.122, 2016.

Michels M, Matte U, Fraga LR, Mancuso ACB, Ligabue Braun R, Berneira EFR, Siebert M, Sanseverino MTV. Determining the pathogenicity of CFTR missense variants: Multiple comparisons of *in silico* predictors and variant annotation databases. Genetics and Molecular Biology, v.42, n.3, p 560-570, 2019.

Montenegro LR, Lerario AM, Nishi MY, Jorge AAL, Mendonca BB. Performance of mutation pathogenicity prediction tools on missense variants associated with 46, XY differences of sex development. Clinics, v.76, p.1-5, 2021.

Pruitt KD, Brow GR, Hiatt SM, Thibaud F, Astashyn A, Ermolaeva O *et al*. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res, v.42, p.756–63, 2014.

Reeb J, Wirth T, Rost B. Variant effect predictions capture some aspects of deep mutational sacnning experiments. BMC bioinformatics, v.21, n.107, 2020.

Richards S, Aziz N, Bale S, Bick D, DAS, S, Gastier-Foster, J. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. Genet Med, v.17, n.5, p. 405–424, 2015.

Sadeghi S, Rafat SA, Alijani A. Evaluation of imputed genomic data in discrete traits using Random forest and bayesian threshold methods. Acta Sci.,Anim. Sic, v.40, 2018.

Sarmento FJQ. Desenvolvimento de uma plataforma de bioinformática integrada aplicada a identificação molecular de microorganismos patogênicos. 2013. Tese (Doutorado em Biotecnologia) - Universidade Federal da Paraíba, João Pessoa.

Scott CJ, Schoeman M, Urban MF. Cancer genetics: an approach to suspected hereditary breast or colorectal cancer. S Afr Med J. v.109, n.4, p.214-218, 2019.

Shihab H A, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics, v.31, n. 10, p.1536–1543, 2015.

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, Day INM, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using Hidden Markov Models. Human Mutation, v.34, n.1, p.57-65, 2013.

Sigrist MS. Mapeamento associativo de locos relacionados á produtividade de grãos em soja. 2012. Tese (Doutorado em genética e melhoramento em plantas), Universidade de São Paulo, Piracicaba: SP.

Sousa JF, Serafim RB, Freitas LM, Fontana CR, Valente V. DNA repair genes in astrocytoma tumorigenesis, progression and therapy resistance. Genetics and Molecular Biology, v.43, n.1, p.1-15, 2019.

Tarazona-Santos E, Kehdy F, Magalhães WCS, Rodrigues MR, Scliar M, Zolini C, Barreto M, Horta B, Pereira AC, Costa MFL. Brasil e a idiossincrasia da miscigenação. Rev. UFMG, Belo Horizonte, v.22, n.1-2, p.232-249, jan. /dez. 2015.

Valverde O, Cezar HXL, Filho VC, Lukesch A, Morais JM. Estudo regional da zona da mata de Minas Gerais. Revista Brasileira de Geografia, v.1, n.1, p.1-139, jan/mar.1958.

Vaser R, Adusumalli S, Leng SN, Sikic M, NG P C. SIFT missense predictions for genomes. Nature protocols, v. 11, n.1, p. 1073-1081, 2016.

Viana NI. Correlação entre polimorfismos genéticos relacionados á hereditariedade, fatores humorais e câncer de próstata. 2017. Tese (Doutorado em Ciências) – Faculdade de Medicina da Universidade de São Paulo, SP.

Wilson C M, LI K, YU X, Kuan P, Wang X. Open Access Multiple-kernel learning for genomic data mining and prediction. BMC Bioinformatics, v.20, n.426, p.1-7, 2019.

Zhang X, Minikel EV, Luria AHOD, Macarthur DG, Ware JS, Weisburd B. ClinVar data parsing. Wellcome Open Research Res, v.2, n.33, 2017.

ISSN: 2179-4804
DOI: https://doi.org/10.20873/jbb.uft.cemaf.v9n2.grosso