

Atividade 6 - Abordagens integrativas para o estudo de condições genéticas

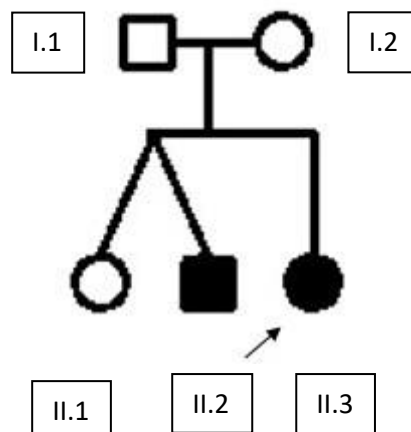
Autor: Camila Cristina Avila Martins, 2020.

Imagine que você acabou de entrar para um serviço de diagnóstico e aconselhamento genético e, em seu primeiro dia, descobre que ficou encarregado de determinar o diagnóstico genético do caso a seguir:

Uma criança (II.3) foi encaminhada ao serviço de aconselhamento genético após receber o diagnóstico clínico de perda auditiva profunda congênita do tipo neurosensorial. Seus pais não apresentavam perda auditiva, assim como não possuíam parentes com a característica anteriormente. Foram realizados alguns testes moleculares de rotina para surdez, entre eles o sequenciamento do gene *GJB2*, para o qual detectou-se que essa probanda possuía uma variante em heterozigose. Defeitos no gene *GJB2*, que codifica a Conexina 26 (expressa na orelha interna), estão associados com perda auditiva sendo que, uma variante em específico, a *35delG*, é responsável por cerca de 70% dos casos de surdez de herança autossômica recessiva. Com os achados moleculares incompatíveis com o desenvolvimento do fenótipo na menina, passou-se a especular que talvez a surdez pudesse ter origem em fatores ambientais ou que outros genes pudessem estar envolvidos. Não havendo mais casos na família, foi informado ao casal que, possivelmente, se tratava de uma mutação *de novo*, nova na menina e não herdada pelos pais, e que, nesse caso, a probabilidade de o casal vir a ter outro filho afetado era semelhante à de qualquer outro casal na população.

Pouco tempo depois, com o nascimento dos gêmeos dizigóticos, o casal voltou ao serviço de aconselhamento após receber o diagnóstico clínico de que o filho II.2 também apresentava o mesmo fenótipo que a irmã mais velha.

A seguir está o heredograma da família:



Heredograma do caso. A seta identifica a probanda, o indivíduo no qual se iniciou o estudo familiar da característica genética.

1. O que você pensa a respeito da primeira informação fornecida ao casal em relação a probabilidade de vir a ter outro filho com perda auditiva?
2. E quanto ao padrão de herança apresentado pela família?

CONHECENDO O OMIM

Provavelmente você deve estar se perguntando como pode começar a estudar o caso dessa família.

O primeiro passo é tentar entender um pouco mais sobre a característica de interesse. Qual gene já foi associado com o fenótipo na literatura? Há mais de um? Qual o padrão de herança da característica?

Uma ferramenta que pode ser utilizada quando não se sabe muito sobre uma determinada característica genética é o OMIM (Online Mendelian Inheritance in Man). Trata-se de um catálogo online que reúne informações sobre genes humanos e fenótipos genéticos.

Entre no site do OMIM: <https://www.omim.org/>

Observe que no mecanismo de busca é possível procurar por características clínicas, fenótipos, genes e mais. Vamos começar procurando por surdez (Deafness). Observe o número de resultados.

Se tudo deu certo (e se você é uma pessoa ansiosa como eu) provavelmente deve estar surtando internamente nesse momento com o número elevado de resultados. Mas, antes que você desista no primeiro dia, vamos tirar algumas informações dessa primeira busca:

Observe que várias entradas encontradas possuem um símbolo “#” seguido de um número. Esse são chamados de “termos ou números OMIM” e nada mais são do que identificadores para um fenótipo descrito no OMIM. Por exemplo: para surdez ligada ao X do tipo 2 (*DFNX2*) o termo OMIM é #304400. (Dica: guarde essa informação para trabalhos, TCCs, teses, dissertações, papers e afins. Caso precise falar sobre uma condição genética, adicione o termo OMIM da característica em questão ao texto. Dessa forma, mesmo que o leitor não tenha muito conhecimento sobre a característica, pode ter acesso as essas informações além de você demonstrar que a característica já foi descrita.)

Ainda analisando as entradas, verifique que há diferentes termos associados à surdez (autossômico recessivo, autossômico dominante, ligado ao X, sindrômico, herança materna) e que, pelas localizações citogenéticas, diferentes genes ou loci parecem estar envolvidos (Xq21.1, 21q22.3, 3p13 ...). Isso já nos mostra que a surdez é uma condição heterogênea em vários aspectos. Um deles é em relação ao padrão de herança. Ela apresenta todos os padrões de heranças conhecidos (autossômico dominante e recessivo, ligado X dominante e recessivo, ligado ao Y e mitocondrial), o que difere são os genes/loci envolvidos em cada um desses padrões.

Vamos especificar essa busca: procure por surdez autossômica recessiva (*deafness autosomal recessive*). Continuamos tendo muitos resultados.

Clique na primeira entrada. Verifique que o OMIM fornece diversas informações sobre esse fenótipo. Em “*Phenotype-Gene Relationships*” é possível verificar a localização citogenética, os números MIMs para o fenótipo e para o gene, o padrão de herança e o gene/loci associados ao fenótipo. Há vários tópicos há serem explorados: “*Description*”, “*Clinical Features*”, “*Mapping*”, “*Molecular Genetics*”, “*Animal Model*”, as referências utilizadas e quem contribuiu com a anotação dessa parte do catálogo. Uma coisa importante a destacar é que o OMIM não possui todas as informações sobre a característica, mas é um bom começo. Em geral, ele tende a conter as primeiras informações de associação de um gene a um fenótipo, como, por exemplo, quando o fenótipo foi descrito pela primeira vez e por quem.

Vá para o topo da página e clique em “*Phenotypic Series*”. Uma tabela contendo todos os fenótipos OMIM para surdez autossômica recessiva deverá aparecer. Na coluna “*Gene/Locus*” observe que diferentes genes já foram associados à surdez autossômica recessiva.

3. Qual era mesmo o nome para isso? Heterogeneidade _____.

Agora que sabemos que diferentes genes podem estar associados à surdez e possuímos uma lista inicial de genes em que podemos procurar (partindo do pressuposto de que a segregação do fenótipo na família siga o padrão de herança autossômico recessivo), vamos para o próximo passo.

ANÁLISE DO EXOMA

Uma vez que os indivíduos afetados já haviam sido testados para genes de rotina para surdez (genes frequentemente mutados em indivíduos com perda auditiva), uma boa opção é tentar utilizar algum método em que possamos explorar mais regiões do genoma ao mesmo tempo. Faremos isso utilizando o sequenciamento massivo paralelo do exoma.

Aqui vão algumas perguntas:

- 4. Seria possível fazer o sequenciamento de um painel de genes nesse caso? E do genoma completo? Qual a vantagem e desvantagem desses em relação ao exoma nesse caso?**
- 5. Se estivéssemos pensando em variantes de número de cópias, esse método seria indicado?**

Você deu sorte e, felizmente, alguém já havia feito o sequenciamento do exoma da paciente II.3. Assim, cabe agora a você analisar os resultados desse exoma. Um dos resultados gerados por essa metodologia é um arquivo no formato VCF (*variant call format*). Abra esse arquivo com o Excel. Ele contém informações sobre as variantes encontradas e anotadas no indivíduo sequenciado (lembrando que o termo variante se refere a uma diferença que o indivíduo pode ter em sua sequência de DNA em comparação com o genoma de referência humano em um locus específico de seu genoma).

Cada linha da tabela indica uma variante sequenciada e, nas colunas, temos diferentes anotações para cada uma das variantes. Observe as 5 primeiras colunas: a primeira coluna (Chr) refere-se ao cromossomo em que a variante foi detectada, a segunda (Start) e terceira (End) colunas à posição, a quinta (Ref) ao nucleotídeo presente nessa posição no genoma de referência e a sexta coluna (Alt) ao nucleotídeo presente nessa posição no indivíduo sequenciado. Assim, para a primeira variante dessa tabela, temos que ela está localizada no cromossomo 1, começando e terminando em 12670, e trata-se de uma alteração de G para C nessa posição.

Vocês devem ter percebido que a quantidade de variantes presentes no arquivo é bem extensa. E não era para menos, afinal, estamos falando do sequenciamento de todos os exons de um indivíduo. Para chegar nas variantes de interesse para o nosso caso, precisamos usar as anotações sobre as variantes a nosso favor e filtrar apenas aquelas que tem maior probabilidade de estarem envolvidas com a perda auditiva.

A primeira coisa que você precisa fazer é filtrar as variantes que estão presentes no indivíduo II.3, isso porque, em um arquivo VCF, além das variantes presentes no seu paciente, há também variantes de outros indivíduos cujas amostras foram sequenciadas na mesma “corrida”. Manter mais de um indivíduo por corrida é interessante pois, viabiliza o custo da metodologia, além de que, esses outros indivíduos podem servir de controle para sua análise. Para isso, vá para as últimas colunas da tabela. Observe que há uma coluna identificada com o registro da sua paciente (II.3). Variantes que não estão presentes em II.3 estão identificadas como 0/0, variantes presentes em heterozigose como 0/1 e, em homozigose, 1/1. Vamos selecionar as em heterozigose e homozigose. Na parte superior do Excel, clique em Dados e, depois, em filtro. Deverá aparecer pequenos quadrados com setas na parte inferior direita de cada célula da primeira linha. Clique no quadrado correspondente da célula II.3, depois em filtro de texto > personalizar filtro. Deverá aparecer uma caixa de texto. Nos dois retângulos da esquerda adicione “começa com” e, nos da direita, adicione em um 0/1 e, no outro 1/1. Clique na opção “Ou” e, em seguida, em “ok”.

Observe que adicionando apenas esse filtro, já diminuimos bastante o número de registros (parte inferior esquerda do Excel).

Como sabemos, em geral, doenças genéticas são raras na população, assim, esperamos que as variantes genéticas associadas às essas doenças também sejam raras. Logo, o segundo passo é filtrar por variantes com baixas frequências em bancos de dados de frequências populacionais mundiais. Vamos adotar nesse exercício a frequência de 1%, isto é, filtraremos todas as variantes com frequência igual ou inferior a 0.01. Vamos utilizar três banco de dados estrangeiros e um brasileiro: Exome Aggregation Consortium (ExAC, exac.broadinstitute.org), NHLBI Exome Sequencing Project (ESP, <http://evs.gs.washington.edu/EVS/>),

1000 Genomes Project (www.internationalgenome.org) e Arquivo Brasileiro Online de Mutações (ABraOM, <http://abraom.ib.usp.br/>).

Utilize o filtro por texto, como mostrado anteriormente, para filtrar variantes iguais ou inferiores a 0.01 ou variante com NA nas colunas "ExAC_ALL", "esp6500siv2_all" e "1000g2015aug_all". Para o banco de dados ABraOM, identificado na tabela como "60mais609_AC", vamos filtrar pelo o número de alelos (não a frequência deles em si). Esse banco contém 609 indivíduos, ou seja, 1218 alelos para cada variante, assim, se consideramos 1%, cerca de 12,18 alelos poderão apresentar a variante rara. Assim, filtraremos por valores iguais ou menores do que 12. Observe que, após adicionar esses filtros eliminamos muitas variantes.

6. **Pensando um pouco sobre o que foi feito agora, você acredita que utilizar o filtro de frequência de variantes em bancos de dados a 1% serve para todos os casos? Por exemplo, se estivéssemos estudando outra condição, como a hipertensão, utilizar esse parâmetro de corte seria sensato? Em que condições isso não seria sensato?**
7. **Qual você acredita que seja a importância de utilizar um banco de frequências populacionais brasileiro?**

Agora que você já conhece o OMIM, que tal utilizá-lo para procurar por variantes presentes em genes anteriormente descritos como relacionados ao desenvolvimento da perda auditiva? Na aba "OMIM_dis", filtre pela palavra "deafness".

Se tudo foi feito corretamente, você provavelmente deve estar vendo uma luz no fim do túnel nesse momento. Com menos variantes, conseguimos analisar melhor uma a uma e verificar quais delas podem estar por trás do fenótipo que observamos.

8. **Há um ponto importante a se considerar quando utilizamos dados do OMIM como filtro. Você imagina qual seria? (dica: no OMIM há informação de condições já descritas).**

Observe a coluna "Gene.refGene". Ela contém a anotação de em que genes essas variantes estão presentes. Na coluna "ExonicFunc.refGene" há a descrição do tipo de mutação (não sinônima, sinônima etc.). Já a coluna "AAChange.refGene" contém a identificação da variante.

Note como essa identificação é feita. Ela contém o nome do Gene, o NM_ (registro do transcrito de acordo com o GenBank), o exon do gene em que a variante está presente, a posição em relação a sequência de código e sequência da proteína em que a mudança aconteceu e qual base e aminoácido foram alterados. Por exemplo, para PAX3:NM_000438:exon4:c.A643G:p.S215G, temos: a variante está localizada no gene PAX3, no exon 4 do transcrito NM_000438. Utilizando como base a sequência de código (c.) (que é aquela contada a partir do primeiro ATG e não contém as sequências 5' e 3' UTRs), houve uma alteração da base adenina na posição 643 para uma guanina, que acarretou na mudança do aminoácido serina na posição 215 da proteína para o aminoácido glicina.

9. **Qual a importância de informar o NM_? Em que implica haver transcritos diferentes para um mesmo gene?**

Embora cada indivíduo normalmente possua cerca de 5 milhões de variantes em seu genoma, quase todas as variantes ou não têm nenhum efeito no portador, ou contribuem para as diferenças naturais entre as pessoas. Como então saber se as variantes que selecionamos são ou não patogênicas?

Nessa parte é interessante olharmos para o InterVar. O InterVar é software de bioinformática utilizado para interpretação clínica de variantes genéticas com base na diretriz ACMG/AMP2015 (*American College of Medical Genetics and Genomics*). Por meio dessa ferramenta, é possível classificar as variantes observadas em benignas, provavelmente benignas, de significância incerta, provavelmente patogênica e patogênica. As diretrizes criadas pela ACMG levam em consideração vários fatores relacionados a essa variante para classificá-la. Dentre eles estão os resultados da predição de patogenicidade *in silico* por algoritmos (programas como *PolyPhen*, *SIFT* e *MutationTaster*), baixas frequências em bancos de dados populacionais, suporte de estudos funcionais que demonstrem que modificações no gene de interesse

alteram o fenótipo como esperado, a segregação da variante na família compatível com a segregação do fenótipo, entre outros.

É importante ressaltar que a classificação gerada pelo InterVar não é absoluta, isto é, ela pode ser modificada conforme mais fatores são considerados. Uma variante inicialmente de significado incerto, por exemplo, pode passar para provavelmente patogênica caso seja comprovado segregação dessa na família compatível com a segregação do fenótipo. Assim, o InterVar pode servir como um guia inicial para selecionar variantes de interesse, mas podemos procurar por mais fatores que possam alterar essa classificação.

10. Olhando para a coluna do InterVar, quais variantes parecem interessantes? Elas estão presentes em homozigose ou heterozigose na paciente?

Parabéns. Você já tem os primeiros genes candidatos para testar nessa família.

CONHECENDO O VARSOME

Que tal conferirmos o que os programas preditores de patogenicidade geram sobre essas variantes? Ao invés de utilizarmos os próprios portais dessas ferramentas, vamos utilizar agora o VarSome.

O VarSome é um portal que reúne informações sobre variantes, mas, mais do que isso, trata-se de um projeto a nível global com o objetivo de compartilhar informações sobre variantes humanas. Além de conter referências cruzadas com muitas outras plataformas e bancos de dados biológicos, também funciona como uma comunidade. Um indivíduo que acesse esse portal pode não somente procurar pela sua variante de interesse como também vincular uma publicação a essa variante, classificar uma variante (segundo critérios ACMG) ou fornecer um comentário sobre uma variante de seu interesse. Assim, é possível compartilhar o seu conhecimento sobre uma variante e se beneficiar da contribuição de outras pessoas.

Entre no site do VarSome: <https://varsome.com/>

Observe que, no mecanismo de busca, é possível inserir o endereço da variante, gene, publicação, fenótipo, identificação no transcrito e afins. Vamos utilizar a identificação da variante no *dbSNP*, o chamado “rs”, para procurá-la. Para isso, verifique o “rs” das variantes candidatas na coluna “*avsnp150*” do arquivo *VCF*. Adicionei o rs que você quer procurar e clique em search.

Fazendo um adendo: O *dbSNP* (do inglês *Single Nucleotide Polymorphism Database*) é um banco de dados público para variações genéticas em uma mesma espécie e entre diferentes espécies. As informações contidas nesse banco de dados são hospedadas pelo *NCBI* (*National Center for Biotechnology Information*) em colaboração com o *NHGRI* (*National Human Genome Research Institute*). Assim como o número *OMIM* e o *NM_*, são registros para fenótipos e transcritos no *OMIM* e *GenBank* (*NCBI*), respectivamente, o “rs” é o registro da variante no *dbSNP*. Apesar do que o nome desse banco de dados dá a entender, o *dbSNP* não apenas hospeda apenas dados de *SNPs*, como também outros tipos de variantes como pequenas deleções e inserções (*indels*), microssatélites (que são pequenas sequências repetitivas em tandem (*STRs*)), entre outros.

Observe a página que abriu. Ela contém várias sessões, com destaque para as sessões:

- “*General Information*”, em que há informações gerais sobre a variante;
- “*Genes*”, em que encontrará o gene em que a variante está localizada. Nessa sessão também há um link que direciona para uma página contendo informações sobre todas as variantes já descritas na plataforma para esse gene;

- “Community Contribution”, em que é possível visualizar informações de outros usuários em relação a variante e contribuir com a plataforma adicionando sua própria classificação para a variante;
- “Publications”, em que você pode verificar se há publicações que mencionem a variante que você buscou (algo maravilhoso, diga-se de passagem);
- “Transcripts”, em que há informações sobre a tua variante em relação aos transcritos. São usados dois registros: um em relação ao *RefSeq* do *NCBI*, nosso conhecido *NM_*; e outro em relação ao *Ensembl*, um banco de dados genômicos do *EMBL (European Molecular Biology Laboratory)*, o registro *ENST*;
- “Region browser”, em que você encontra informações sobre a sequência de bases próximas a da variante buscada. Isso inclui: quais os scores de conservação de cada base na região, se há variantes próximas a que você buscou, suas patogenicidades e quais as frequências delas em alguns bancos de dados;
- “Germline Classification”, em que há a classificação da variante segundo os critérios da *ACMG* (note que é possível marcar e desmarcar critérios dependendo do que você observa no caso em que está estudando). É possível saber quais critérios estão sendo utilizados. Clique em “?” para saber mais sobre os critérios;
- “ClinVar”, em que se observa o resumo das informações fornecidas pela base de dados ClinVar, conectando variantes com fenótipos já associados a elas;
- “Frequencies”, que mostra as frequências para a variante buscada no banco de dados *gnomAD (Genome Aggregation Database)*. Algo interessante sobre isso é que, o *gnomAD exomes*, é o banco de dados sucessor do *ExAC* que vimos anteriormente;
- “Conservation scores”, na qual há scores quanto ao nível de conservação da região em que a variante buscada está presente gerados por diferentes ferramentas;
- “In-Silico Predictors”, que contém dados de predição de patogenicidade da variante buscada gerados por diferentes programas (aqui entram programas como o *SIFT*, *Polyphen* e *MutationTaster* mencionados anteriormente);
- “Structural Variants”, em que é possível visualizar transcritos presentes na região em que sua variante está e verificar se há variantes estruturais nessa região;
- “Protein Viewer”, em que é possível visualizar a estrutura da proteína formada pelo gene contendo a variante buscada. Além disso, o visualizador permite verificar os resíduos para os quais já foram encontradas variantes e as classifica segundo a patogenicidade.

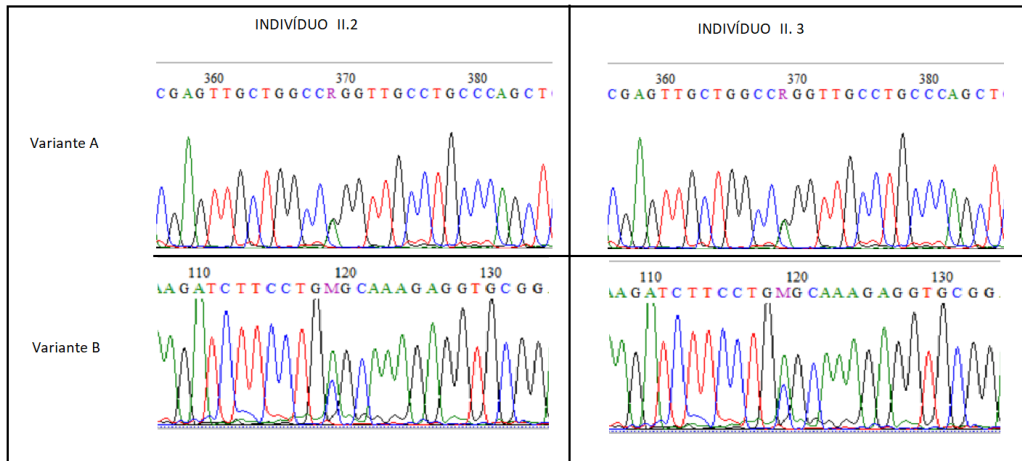
Utilizando o Varsome, verifique a predição de patogenicidade de todas as variantes de interesse encontradas no exoma.

11. Quais foram as predições de patogenicidade encontradas para cada uma delas?

12. Considerando essas predições, quais variantes ainda parecem interessantes?

SEQUENCIAMENTO DE SANGER

Você resolveu então realizar o sequenciamento de Sanger dessas variantes candidatas em ambos filhos dessa família (II.2 e II.3) para confirmar a segregação dessas nos indivíduos afetados. Os eletroferogramas resultantes desse sequenciamento estão identificados a seguir.



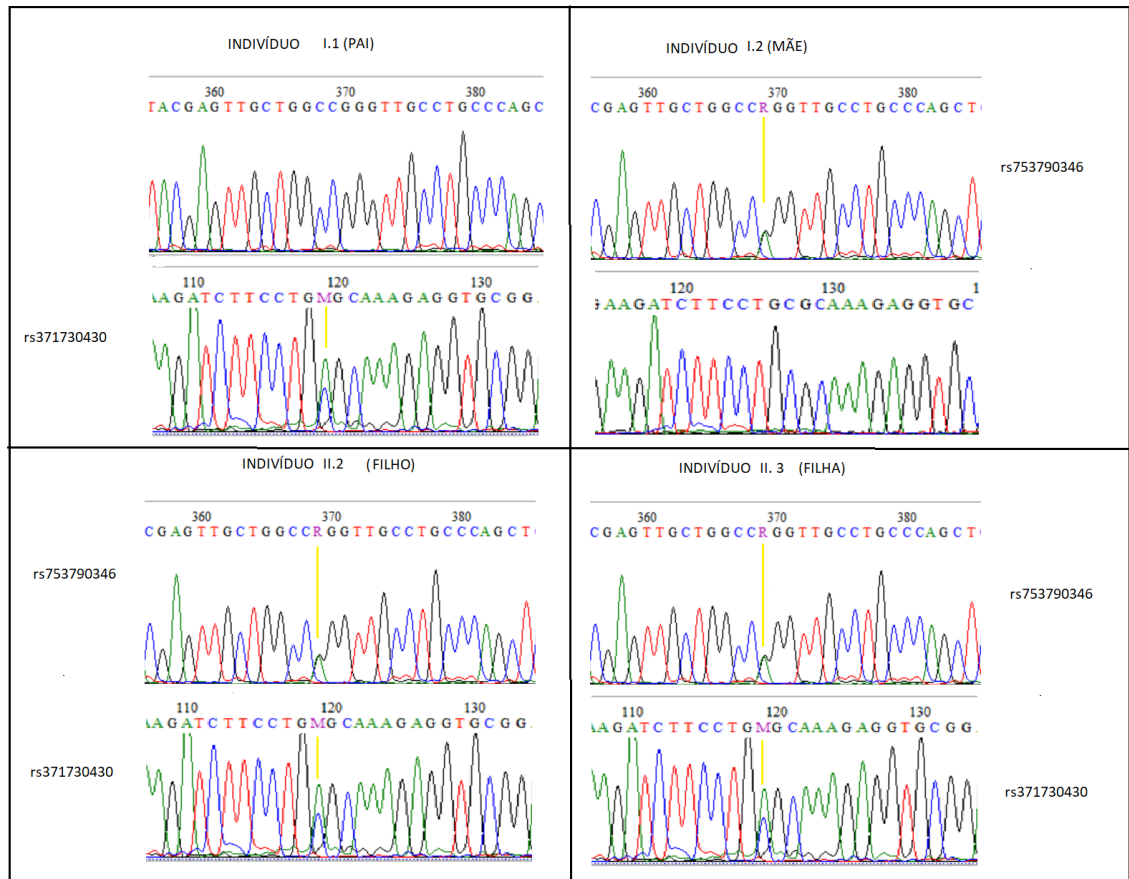
13. Qual variante é a MYO15A:NM_016239:exon10:c.G4252A:p.G1418R, rs753790346? A ou B? (Utilize a tabela a seguir para resolver esse item).

Tabela contendo a representação padrão da IUPAC

Símbolo	Bases representadas
R	A, G
Y	C, T
M	A, C
K	G, T
S	C, G
W	A, T
H	A, C, T
B	C, G, T
V	A, C, G
D	A, G, T
N	A, C, G, T

14. E a MYO15A:NM_016239:exon42:c.C8080A:p.R2694S, rs371730430?
15. Olhando para todos esses resultados, qual a sua hipótese para explicar a condição de perda auditiva nessa família? (lembre-se do padrão de herança já descrito para os genes afetados).
16. Considerando essa hipótese, podemos fechar o caso dessa família somente com os sequenciamentos desses genes nos filhos afetados?

Para observar a segregação da variante na família você também testou os pais. O resultado do sequenciamento de Sanger para as variantes candidatas é mostrado a seguir:



Eletroforese identificando a segregação das variantes na família. Observe que o pai apresenta uma variante, enquanto, a mãe, a outra. Os filhos afetados possuem ambas variantes em heterozigose composta.

Parabéns. Você acaba de resolver seu primeiro caso de diagnóstico genético.

REFERÊNCIAS

- “OMIM - Online Mendelian Inheritance in Man.” n.d. Accessed October 5, 2020. <https://www.omim.org/>.
- Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. “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.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24.
- “Varsome The Human Genomics Community.” 2015. November 25, 2015. <https://varsome.com/>.

CONTRIBUIÇÕES

- Larissa Nascimento, aluna de mestrado do departamento de Genética e Biologia do Instituto de Biociências. O caso utilizado para essa atividade faz parte da casuística de seu trabalho de mestrado.
- Professores Dra. Regina Célia Mingroni Netto, Dra. Luciana Amaral Haddad, Dr. Michel Satya Naslavsky e Dr. Eduardo Gorab por cederem o “Roteiro de Aulas Práticas Análises Computacionais de Sequências de ácidos Nucléicos”, da disciplina de Biologia Molecular para Bacharelado – BIO307 que serviu de inspiração para a criação dessa atividade.

