UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE BIOCIÊNCIAS
HOSPITAL DE CLÍNICAS DE PORTO ALEGRE
SERVIÇO DE GENÉTICA MÉDICA

MALU BETTIO SOARES

ANÁLISE DO GENE *GNPTAB*: CONSTRUÇÃO DE HAPLÓTIPOS E ORIGEM DA MUTAÇÃO MAIS COMUM EM PACIENTES COM MUCOLIPIDOSES II e III

Porto Alegre
2018

MALU BETTIO SOARES

ANÁLISE DO GENE *GNPTAB*: CONSTRUÇÃO DE HAPLÓTIPOS E ORIGEM DA MUTAÇÃO MAIS COMUM EM PACIENTES COM MUCOLIPIDOSES II e III

Trabalho de Conclusão de Curso apresentado ao Instituto de Biociências, da Universidade Federal do Rio Grande do Sul, como requisito à obtenção do grau de Bacharel em Ciências Biológicas.

Orientadora: Profa. Dra. Ida Vanessa Doederlein Schwartz

Co-orientadora: Dra. Fernanda Sperb Ludwig

Hospital de Clínicas de Porto Alegre

Serviço de Genética Médica

Porto Alegre
2018

MALU BETTIO SOARES

ANÁLISE DO GENE *GNPTAB*: CONSTRUÇÃO DE HAPLÓTIPOS E ORIGEM DA MUTAÇÃO MAIS COMUM EM PACIENTES COM MUCOLIPIDOSES II e III

Trabalho de Conclusão de Curso apresentado ao Instituto de Biociências, da Universidade Federal do Rio Grande do Sul, como requisito à obtenção do grau de Bacharel em Ciências Biológicas.

BANCA EXAMINADORA:

_____

Andreia Turchetto-Zollet (UFRGS)

_____

Taciane Borsatto (UFRGS-HCPA)

Porto Alegre
2018

AGRADECIMENTOS

Agradeço à Fernanda Sperb-Ludwig, minha co-orientadora, pelos ensinamentos, pela paciência, dedicação e disponibilidade na realização deste trabalho, além do incentivo, especialmente em momentos em que eu pensava em desistir. Também à minha orientadora, Ida Vanessa Doederlein Schwartz pelos ensinamentos, incentivos e pela oportunidade de bolsa e permanência no laboratório BRAIN. A todos os colegas do BRAIN por sempre se mostrarem dispostos a ajudar, acrescentar e incentivar o meu crescimento e de todos os colegas. Às minhas duas avaliadoras da banca, Taciane Borsatto e Andréia Carina Turchetto-Zolet, por ajudarem também no meu crescimento profissional.

Aos meus amigos do Clube (Amanda, Bruna, Débora, as Jennis, Lilith, Lucas, Natália, Pâmela, Patrícia, Pietro, Samuel e Verônica), que estiveram comigo esse tempo todo, apoiando durante tempos de crise e me dando força. Em especial à minha colega de laboratório e amiga Cristal pela parceria, por ouvir minhas lamentações e ajudar muito nos problemas com o TCC, sempre alegrando meus dias.

O principal agradecimento é à minha família, pelo apoio e compreensão, tentando de todo o jeito proporcionar tranquilidade nesses tempos de estresse. Principalmente à minha mãe, Valéria, que além de me dar a vida e ser a minha melhor amiga, sempre me inspirou por ser a pessoa mais fantástica que eu conheço e me incentivou a ser uma pessoa melhor, além de ter ajudado a revisar e montar o TCC. Aos meus irmãozinhos cães Fred e Luna, pela companhia nas noites em claro e pelo amor incondicional sempre. À tia e Dinda, Marisa, que é minha segunda mãe, e ao primo/irmão, João Pedro, além do meu tio José Luis, tia Rosemeri, às primas Ellen, Aline, Dani, que sempre torcem muito por mim. Vocês são minha vida, não consigo nem descrever a sorte e bênção que é ter vocês como família. Também gostaria de agradecer ao meu amor, Fernando Martins, por ter

estado comigo todo esse tempo aturando o estresse, tentando sempre me motivar, acalmar, alegrar, dar espaço e tranquilidade para concluir o TCC.

SUMÁRIO

1 INTRODUÇÃO

Erros Inatos do Metabolismo (EIM) são distúrbios de natureza genética que acarretam anormalidades na síntese ou catabolismo de proteínas, carboidratos ou lipídios através de defeito enzimático ou transporte de proteínas, resultando em bloqueio ou prejuízo do funcionamento de vias metabólicas (Rao et al., 2009; El-Husny, Fernandes-Caldato, 2006). São conhecidas mais de 500 doenças associadas aos EIM, representando cerca de 10% de todas as doenças genéticas (Sanseverino et al, 2000).

Entre os EIM estão as doenças lisossômicas (DL), um grupo heterogêneo composto de mais de 50 doenças genéticas distintas, caracterizadas por progressivo acúmulo de substratos específicos nos lisossomos. As DLs são a causa mais comum de doença neurodegenerativa pediátrica (Coutinho; Alves; 2016). Embora as primeiras descrições clínicas de pacientes com DLs tenham sido feitas em 1881 por Warren Tay, apenas cerca de 50 anos depois se descobriu a natureza bioquímica de alguns dos produtos acumulados. Só depois, em 1963, Hers finalmente demonstrou uma correlação entre defeito enzimático e problemas de acúmulo de substrato intralisossomal (Filocamo; Morrone, 2011; Coutinho; Alves; 2016). Esse acúmulo de substrato, parcialmente ou não digerido, pode ocasionar uma cascata de efeitos patogênicos, resultando em complexos quadros clínicos caracterizados por envolvimento multissistêmico, que ocorre devido à deficiência de enzimas, proteínas lisossômicas e, em alguns casos, até proteínas não lisossômicas, mas envolvidas na biogênese lisossômica. A maioria das proteínas relacionadas com essas doenças reside no lúmen do lissosomo e uma minoria são proteínas de membrana (Kingma et al, 2015; Filocamo; Morrone, 2011; Ballabio; Gieselmann, 2009; Coutinho; Alves, 2016).

As DLs normalmente são classificadas quanto ao substrato predominantemente acumulado, o que é uma classificação considerada útil e muito bem aceita clinicamente. Entretanto, na maioria das DLs mais de um composto é acumulado e em outras os substratos acumulados podem ser bastante heterogêneos (Ballabio; Gieselmann, 2009). A maioria delas apresenta um padrão de herança

autossômico recessivo sendo normalmente monogênicas e, para um grande número delas, numerosas mutações foram descritas no mesmo gene. Algumas dessas mutações levam à total perda de atividade enzimática, enquanto outras levam à atividade reduzida. No entanto, não foi encontrada ainda para várias dessas doenças uma correlação clara entre genótipo e fenótipo. (Filocamo; Morrone, 2011; Futerman; van Meer, 2004).

Embora as DLs sejam doenças consideradas raras isoladamente, quando analisadas em conjunto, apresentam frequência relevante: sua incidência foi estimada em 1:7700 nascimentos (Meikle et al, 1999). Em países com alta taxa de consanguinidade, pode haver uma incidência maior de distúrbios hereditários (Moammar et al, 2010). As manifestações clínicas variam de leves a graves, podendo não ser evidentes logo no nascimento na maioria dos casos, aparecendo geralmente na infância. Além disso, os métodos de diagnóstico são relativamente sofisticados, o que dificulta o mesmo em países em desenvolvimento e implica uma incidência provavelmente subestimada dessas doenças (Giugliani, 2017).

1.1 LISOSSOMOS E ROTA ENDOSSOMO-LISOSSOMAL

Os lisossomos são organelas citoplasmáticas responsáveis pela degradação de vários tipos de macromoléculas. Embora sejam organelas consideradas digestivas, as hidrolases ácidas e proteínas associadas que compõem a membrana lisossômica são relativamente duradouras, enquanto os constituintes endógenos são continuamente substituídos por compostos mais recentemente sintetizados (Kornfeld, 1989).

Os lisossomos são constituídos de uma membrana externa limitante e vesículas intralisossomais (Sandhoff; Kolter, 1996; Futerman; van Meer, 2004). Atualmente não são mais consideradas apenas organelas digestivas ou o destino final dos compostos degradados, mas organelas centrais para a homeostase celular metabólica, coordenando uma rede complexa e interativa de organelas intracelulares. O lisossomo possui funções específicas e envolvidas integralmente na fagocitose,

autofagia, exocitose, regulação de receptores, sinalização intracelular, imunidade, pigmentação e neurotransmissão; o aprofundamento dos estudos acerca das doenças que derivam de falhas na biogênese desta organela pode esclarecer muitos mecanismos celulares ainda desconhecidos (Coutinho; Ales; 2016; Futerman; van Meer, 2004).

Os lisossomos fazem parte da rota endossomo-lisossomal, juntamente com endossomos precoces, tardios e corpos multivesiculares – também podendo ser chamados de compartimentos lisossomais – e constituem até 5% do volume intracelular das células animais (Hu et al, 2015).

O compartimento lisossômico como um todo consiste de uma coleção de vacúolos de composição, morfologia, localização e densidade heterogênea. Essa distinção se dá pelas diferenças de nível de degradação dos substratos dentro dos vacúolos individuais, além de elementos de fusão entre os vacúolos (Huotari; Helenius, 2011). Essa característica reflete a maneira pela qual os lisossomos são formados e contrasta com a relativa uniformidade das outras organelas celulares, sendo diferenciados de endossomos tardios pela ausência de receptores de Manose-6-fosfato (M6P) (Luzio; Pryor; Bright; 2007; Alberts et al, 2014). Já os endossomos precoces não têm tantas semelhanças com a estrutura final do lisossomo, inclusive posicionando-se na periferia da célula, sendo o local onde a maioria das moléculas internalizadas são mantidas (Hu, et al, 2015).

Os lisossomos possuem um ambiente intracelular ácido com pH de até 4.5, distinto do pH citoplasmático que é de cerca de 7.0-7.2, o que só é possível pela existência da bicamada lipídica limitante que os envolve, possibilitando também que haja condições ótimas para funcionalidade das hidrolases e de outras enzimas que ali estão localizadas. Esse ambiente ácido é formado por uma queda rápida progressiva – devido à acidificação vacuolar – no pH luminal dos endossomos precoces, quando transformados em endossomos tardios e posteriormente fusionados com os lisossomos na rota endossomo-lisossomal, ocorrendo concomitantemente ao aumento da concentração das hidrolases ácidas e à migração para o centro da célula. A formação de corpos multivesiculares é resultante desse processo, representando a transformação dos endossomos precoces em tardios (Hu et al, 2015).

Além disso, todas as interações celulares nas quais os lisossomos estão envolvidos são mediadas pela membrana e representam o principal mecanismo pelo qual os substratos são internalizados ao lisossomo para degradação (Settembre; Ballabio, 2014). Os produtos dessa degradação são transportados de volta para o citoplasma por proteínas específicas de transporte localizadas nas membranas (Coutinho; Alves; 2016; Tettamanti et al, 2003).

São conhecidas mais de 60 hidrolases ácidas contidas nos lisossomos: fosfatases, nucleases, glicosidases, proteases, peptidases, sulfatases e lipases, as quais são responsáveis pelo processo de degradação da maioria das macromoléculas e mais de 100 proteínas de membrana, incluindo as que propiciam a funcionalidade das hidrolases lisossômicas (Luzio et al, 2014; Schwanke; Schröder; Saftig, 2013). Também, os lisossomos agem como local de ativação de diversas atividades proteolíticas (Hu et al, 2015). As proteínas transmembranas lisossômicas mais abundantes registradas são as LIMP-1 e LIMP-2, constituindo cerca de 50% de todas as proteínas da membrana (Luzio et al, 2014). A maioria delas é altamente glicosilada, o que auxilia a proteger a membrana das proteases localizadas no lúmen do lisossomo (Alberts et al, 2014).

Os lisossomos podem ser caracterizados como o destino comum para o qual diversas rotas distintas de tráfego intracelular convergem; seu processo de síntese é bastante complexo e ainda não bem entendido, no qual uma rota que leva para fora do Retículo Endoplasmático através do Complexo de Golgi entrega a maioria das hidrolases, enquanto ao menos quatro outras rotas fornecem outros substratos para que sejam digeridos nos lisossomos (Alberts et al, 2014).

As hidrolases ácidas, conforme sintetizadas, são endereçadas aos lisossomos através de elementos de transição, processo que começa a ocorrer no Complexo de Golgi (Rohrer; Kornfeld, 2001). Essas hidrolases recebem marcadores Manose-6-Fosfato (M6P) pela ação de duas enzimas: N-acetilglicosamina-1-fosfotransferase (referida também como GlcNAc-1-fosfotransferase, E.C. 2.7.8.17, codificada pelo gene *GNPTAB* e *GNPTG*) e N-acetilglicosamina-1-fosfodiester alfa-N-acetilglicosaminidase (E.C 3.1.4.45, codificada pelo gene *NAGPA*) logo depois que saem do Retículo Endoplasmático Rugoso, ainda na parte *cis* do complexo de Golgi

e, devido ao reconhecimento do M6P, localizado na ponta das cadeias de oligossacarídeos N-ligados das hidrolases (fig. 1), podem adentrar os endossomos tardios e em seguida os lisossomos (por fusão) através de receptores de M6P (M6PR). Ou seja, a adição e posterior reconhecimento do marcador são processos-chave para o endereçamento correto das hidrolases lisossômicas e, portanto, para o funcionamento correto dos lisossomos (fig. 2; Alberts et al, 2014).

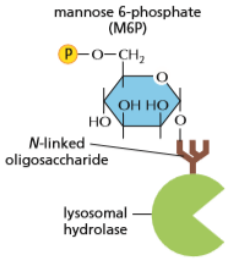

Figura 1: Localização do marcador M6P nas hidrolases lisossômicas
Fonte: Alberts et al, 2014

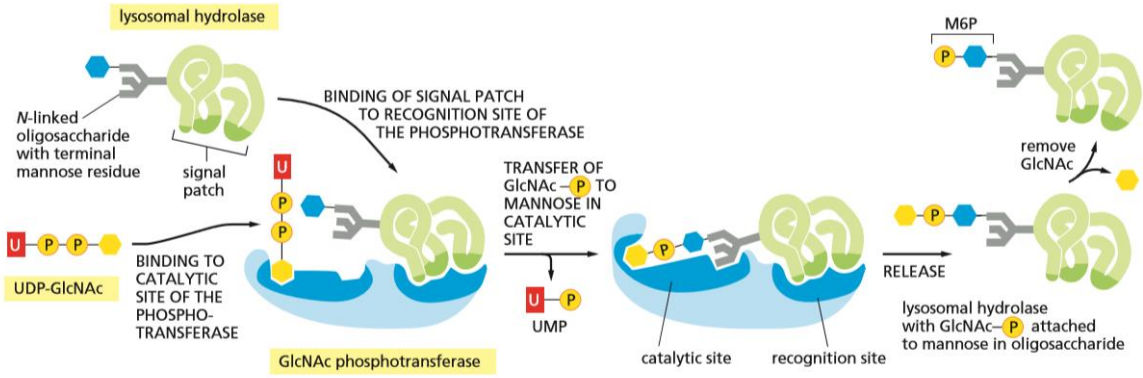

Figura 2: Reconhecimento da uma hidrolase lisossômica.
Fonte: Alberts et al, 2014.

1.2 A N-ACETILGLICOSAMINA-1-FOSFOTRANSFERASE

A identificação e caracterização da enzima N-acetilglicosamina-1-fosfotransferase foi realizada primeiramente por Bao et al. em 1996, através de bovinos, onde foi determinada sua associação à membrana e sua localização no complexo de Golgi, caracterizando-a em um complexo de 540kDa composto de seis subunidades de homodímeros de 166kDa e 51kDa, ligados a dissulfitos e

subunidades não covalentemente associadas de 56kDa (Bao et al, 1996a). Esses achados possibilitaram que propriedades específicas da enzima fossem atribuídas a subunidades específicas, além de determinar que a proteína promove a fosforilação (adição de um grupo fosfato) seletivamente a enzimas lisossômicas e que essa seletividade é uma propriedade da enzima em si e não um fator acessório. A atividade da enzima foi registrada em pH de 5.7 a 9.4, com funcionalidade ótima em pH 6.7 a 7.5, porém, a estabilidade da enzima foi identificada em pH 5 a 11 (Bao et al, 1996b). Só então, no ano de 2005, Tiede et al. caracterizaram a enzima humana com 1256 aminoácidos e com massa molecular de 144kDa. Demonstraram também que são dois domínios transmembrana e 19 potenciais sítios de N-glicosilação. Além disso, a organização modular do gene *GNPTAB* é preservada em proteínas equivalentes em ratos, cães, galinhas e *zebrafish*.

Essa enzima é codificada por um processo complexo e de interação de 2 genes distintos: *GNPTAB* e *GNPTG*. O gene *GNPTAB* tem tamanho de 85kB, é formado por 21 éxons, está localizado no cromossomo 12, na posição 23.2 do braço longo (q) e codifica os precursores de quatro subunidades (duas α e duas β) da enzima GlcNac-1-phosphotrasnferase, sendo as outras duas subunidades γ codificadas pelo gene *GNPTG* que, distintamente do *GNPTAB*, está localizado no cromossomo 16, na posição 13.3 do braço curto (p) e contém 11 éxons. (Raza, et al, 2015; Hashemi-Gorji et al 2016; Cathey et al, 2008; Tiede et al, 2005).

No primeiro passo de formação, o gene *GNPTAB* sintetiza a proteína precursora das duas subunidades α e β, que têm função catalítica. Esse precursor é uma proteína de membrana tipo III de 1256 aminoácidos e tamanho de cerca de 144kDa, com terminações N e C viradas para o citosol (Tiede et al, 2005; Franke; Braulke; Storch, 2013).

A clivagem do precursor ocorre entre os resíduos Lys-928 e Asp-929 e é catalisada pela *"site-1-protease*" (S1P) localizada no complexo de Golgi, o que resulta nas subunidades α e β cataliticamente ativas. Já as subunidades γ são sintetizadas como uma glicoproteína solúvel de 305 aminoácidos e têm como funções a otimização da atividade catalítica das subunidades α e β e auxílio no reconhecimento da proteína determinante das hidrolases (Qian et al, 2013; Tiede et al, 2005).

A principal funcionalidade dessa enzima é no processo de adição do marcador M6P para endereçamento de hidrolases lisossômicas, o que ocorre pela ação conjunta de duas enzimas distintas que se relacionam nessa função: no primeiro passo, a enzima GlcNAc-1-fosfotransferase catalisa a ligação covalente entre N-acetilglicosamina-1-fosfato a partir do UDP(Uridina Difosfato)-N-acetilglicosamina, para grupamentos hidroxil no carbono 6 de resíduos terminais de manose das cadeias de oligossacarídeos das enzimas destinadas aos lisossomos, gerando um fosfodiéster intermediário. No segundo passo, a Enzima Descobridora (*Uncovering Enzyme* – UCE), codificada pelo gene *NAGPA*, remove um grupamento de N-acetilglicosamina, hidrolisando-o e expondo o marcador M6P para que seja reconhecido e, então, as hidrolases são direcionadas aos endolisossomos e finalmente aos lisossomos. O processo inicia-se na porção *cis* do complexo de Golgi e a última parte ocorre na porção *trans* do complexo de Golgi, ou *Trans-Golgi Network* (TGN) (Kang et al, 2010; Kornfeld et al, 1989).

O sucesso da adição do marcador e consequente transporte ao lisossomo ocorre devido à especificidade de ligação do marcador a enzimas lisossômicas. Várias hidrolases já tiveram sequências clonadas e não foi encontrada nenhuma similaridade, indicando que a proteína não reconhece uma sequência específica, mas sim uma funcionalidade específica comum (Kornfeld, 1989; Luzio et al, 2014). Segundo Qian et al (2013), a especificidade da reação de fosforilação de hidrolases ácidas é determinada pela habilidade da enzima N-acetilglicosamina-1-fosfotransferase de reconhecer um determinante de proteína dependente de conformação, presente em hidrolases ácidas e ausente em glicoproteínas não lisossomais.

1.3 MUCOLIPIDOSES II E III ALFA/BETA E MUCOLIPIDOSE III GAMA

As Mucolipidoses (MLs) II e III alfa/beta e ML III gama são doenças lisossômicas, genéticas e raras, com padrão de herança autossômico recessivo, caracterizadas por defeito na enzima GlcNAc-1-fosfotransferase, o que ocasiona

perda total ou parcial da atividade enzimática. O defeito enzimático acaba prejudicando o tráfego das hidrolases lisossômicas, tendo como consequência a secreção excessiva de enzimas lisossômicas (Bargal et al, 2006; Coutinho et al, 2011; Koehne et al, 2016). As deficiências celulares de múltiplas enzimas nos lisossomos levam ao acúmulo de macromoléculas não degradadas como proteínas, lipídios ou glicosaminoglicanos (GAGs), e mal funcionamento dos lisossomos devido à deficiência intracelular das hidrolases ácidas (Koehne et al, 2016; Encarnação et al, 2009).

A ML II alfa/beta é a forma mais severa da doença, com quadro clínico composto de crescimento ósseo prejudicado, face infiltrada, hipertrofia gengival, macroglossia, hérnias inguinais, infiltrações cutâneas, limitação articular, surdez, atraso psicomotor e baixa estatura (Cathey et al, 2010; Koehne et al, 2016; Plante et al, 2008). Os sintomas podem ser detectados até antes do nascimento e, devido à progressão da doença, geralmente ocorre o óbito na primeira década de vida em função de problemas cardiopulmonares (Hashemi-Gorji et al, 2016).

Apresentando características clínicas mais brandas, as ML III alfa/beta e ML III gama apresentam progressão mais lenta que as ML II alfa/beta, ocorrendo o aparecimento dos sintomas nas primeiras décadas de vida. Devido a essa progressão mais lenta, a sobrevivência se estende até a vida adulta, podendo chegar a 8ª década. As primeiras manifestações clínicas da doença geralmente são a rigidez de articulações das mãos e ombros, escoliose, baixa estatura e “mãos em garra” (Ludwig 2016; van Meel; Kornfeld, 2016).

As ML II e III alfa/beta são ocasionadas por mutações no gene *GNPTAB*, e a ML III gama por mutações no gene *GNPTG* (Cathey et al, 2010).

O diagnóstico dessas doenças é feito através de métodos bioquímicos e moleculares. No centro de referência de Erros Inatos do Metabolismo do Hospital de Clínicas de Porto Alegre, o diagnóstico bioquímico é realizado através da medição da atividade das enzimas Arilsulfatase A (ARSA, EC 3.1.6.8), α-l-iduronidase (IDUA; EC 3.2.1.76), iduronate-sulfatase (IDS; EC 3.1.6.12), β-glucuronidase (GUSB; EC 3.2.1.31), e β-hexosaminidase (EC 3.2.1.30) em fibroblastos, assim como a análise de Glicosaminoglicanos (GAGS). A análise molecular é feita através de

sequenciamento dos genes *GNPTAB* e *GNPTG* (Cury et al, 2013). Em outros centros, o diagnóstico é feito através de ensaio enzimático da GlcNAc-1-fosfotransferase com substância radioativa.

Até dezembro de 2016, 160[1] mutações foram identificadas no gene *GNPTAB*, sendo 153 causadoras das MLs II e III alfa/beta e o restante das mutações causadoras de gagueira. Para o gene *GNPTG*, foram identificadas 42[2] mutações patogênicas, 26 causadoras de ML III gama e o restante causadoras de gagueira (Raza et al, 2016). Os estudos de Raza et al (2016) corroboram a informação de que as mutações causadoras de MLs são fundamentalmente distintas das causadoras de gagueira.

Embora exista um número alto de mutações distintas causadoras de MLs e ainda que a maioria seja considerada única ou rara, a microdeleção c.3503_3504delTC (rs34002892) foi encontrada como uma mutação associada a efeito fundador na população franco-canadense, além de ter sido encontrada em variadas populações, registrando uma ampla distribuição geográfica, sendo esta a mutação mais frequente causadora de ML II/III alfa/beta. Essa mutação é caracterizada como uma mutação patogênica de mudança de fase de leitura e está posicionada no éxon 19. Ocasiona, na primeira posição da deleção, a troca do aminoácido Leucina para Glutamina e a consequente troca de toda a fase de leitura e dos aminoácidos ali presentes [3](Coutinho et al, 2011; Plante et al, 2008).

Na população dos Estados Unidos foi estimada uma frequência de 22% de presença da mutação em 61 pacientes de ML II/III alfa/beta (Cathey et al, 2010); Tappino et al (2009) encontrou 51% de 38 pacientes (maioria - 83% - Italianos, 4 Argentinos e 2 Paquistaneses) de MLs com a mutação, e Plante et al (2008) detectou a mutação em 16 pacientes e 9 pais de pacientes em uma população fundadora do Canadá.

Na população latina, a frequência encontrada para a mutação c.3503_3504delTC foi de (0,043%), 5/11540 alelos afetados/alelos analisados segundo o Projeto ExAC. Segundo o ABraOM, banco de dados de variantes

[1] HGMD: The Human Gene Mutation Database. Disponível em: http://www.hgmd.cf.ac.uk/ac/index.php
[2] HGMD: The Human Gene Mutation Database. Disponível em: http://www.hgmd.cf.ac.uk/ac/index.php
[3] National Center for Biotechnology Information – NCBI: dbSNP – The Single Nucleotide Polymorphism database. Disponível em: https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=34002892

genômicas brasileiras, a frequência detectada foi 1/1218 alelos, (0,082%). A frequência alélica total no Projeto ExAC é 0,054%, 66/121350 alelos. Segundo o Projeto 1000 genomas, a frequência total é 119/245997, (0,048%) enquanto a frequência para a população latina (referida como Americana – AMR – pelos autores) é de 17/33541, (0,050%).

Para pacientes brasileiros de ML II/III alfa/beta, a frequência foi determinada no trabalho de Cury et al (2013) como presente em 45% (11/24) dos alelos e em 40% dos alelos (10/25) em Ludwig et al (2016).

O estudo de Coutinho et al (2011) determinou a idade da mutação patogênica c.3503_3504delTC em 2063 anos (+-729 anos) e sua provável origem através da construção de haplótipos.

1.4 HAPLÓTIPOS, SNPS E DESEQUILÍBRIO DE LIGAÇÃO

Os haplótipos por definição são conjuntos de alterações polimórficas herdados em conjunto, e a análise da existência de desequilíbrio de ligação (DL) entre polimorfismos e mutações em amostras de pacientes permitem traçar um histórico populacional de mutações relevantes. O desequilíbrio de ligação refere-se à associação não aleatória entre alelos de *loci* adjacentes (Ahmad et al, 2003), e permite que análises de evolução do gene sejam realizadas. Para tanto, a análise de variantes e marcadores moleculares associados a mutações patogênicas possibilitam traçar também um histórico do gene relacionado, assim como a dispersão de alelos mutantes. Nesse sentido, os SNPs (*Single Nucleotide Polymorphisms*) podem ser uma importante ferramenta de análise. Aproximadamente 90% da variação genética do genoma humano é composta por SNPs, resultado de mutação de ponto – que podem ser substituições ou *indels* (inserções/deleções) – na sequência cromossômica. Recentemente, os SNPs estão sendo cada vez mais utilizados como marcadores moleculares para estudos de história evolutiva de populações e até de especiação. Ao contrário de outros marcadores, como microssatélites, os SNPs têm uma taxa de mutação relativamente baixa, descrita como um intervalo de 10^{-8} a 10^{-9}. Múltiplas mutações no mesmo *loci* em SNPs são extremamente improváveis, então a

maioria dos SNPs são bialélicos, facilitando a genotipagem (Brumfield et al, 2003; Turchetto-Zolet et al, 2017).

2 JUSTIFICATIVA

Embora sejam raras, as ML II e III alfa/beta são doenças graves e que podem culminar em óbito. Devido à progressão da doença, geralmente os pacientes necessitam de cuidados intensivos e em tempo integral por parte da família, o que ocasiona sofrimento. Como há uma baixa incidência mundial, as MLs são ainda relativamente desconhecidas, o que dificulta o encaminhamento devido e o diagnóstico correto.

A importância do conhecimento da diversidade genética das populações, dos haplótipos presentes e análises de desequilíbrio de ligação (DL) se dá pela possibilidade de mapeamento genético, identificação de genes associados a doenças, conhecimento da história demográfica de populações e do funcionamento da ação de forças seletivas. Além disso, tal conhecimento pode facilitar estimativas de incidência, aconselhamento genético e diagnóstico efetivo das MLs.

O desequilíbrio de ligação entre os alelos carregando marcadores como SNPs pode ser indicador das forças seletivas que agem estruturando um genoma. Atualmente, a maioria dos estudos com DL tem como objetivo entender a história evolutiva e eventos demográficos de populações, fazer mapeamento de genes que podem estar associados a caracteres quantitativos e doenças hereditárias, além do entendimento da evolução de haplótipos (Slatkin, 2009).

3 OBJETIVOS

3.1 OBJETIVOS GERAIS

- Caracterizar os haplótipos de pacientes brasileiros com MLs II e III alfa/beta;
- Identificar a origem da mutação c.3503_3504delTC na população brasileira.

3.2 OBJETIVOS ESPECÍFICOS

- Caracterizar os haplótipos formados por 9 polimorfismos genéticos: c.-41_-39delGGC (rs76300806), c.18G>A (rs4764655), c.27G>A (rs222504), c.365+96_365+97delGT (rs4015837), c.365+145C>T (rs2108694), c.1285-166G>A (rs7963747), c.1932A>G (rs10778148), c.3135+5T>C (rs759935) e c.3336-25T>C (rs3736476) e pela mutação patogênica mais frequente em pacientes c.3503_3504delTC, todos localizados no gene *GNPTAB*, em pacientes brasileiros com diagnóstico molecular de ML II e III alfa/beta;

- Determinar os passos mutacionais relacionados à evolução dos haplótipos caracterizados;

- Analisar a existência de desequilíbrio de ligação entre os polimorfismos citados presentes no gene *GNPTAB* e também com a mutação c.3503_3504delTC;

- Determinar a origem da mutação c.3503_3504delTC na população brasileira.

REFERÊNCIAS

Ahmad T, Neville M, Marshall SE, Armuzz A, Mulcahy-Hawes K, Crawshaw J, Welsh KI. Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. Hum Mol Genet 2003;12(6):647-656.

Alberts B, Johnson A, Lewis J, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell. Journal of Chemical Information and Modeling 2014;(1):695-752.

Ballabio A, Gieselmann V. Lysosomal disorders: From storage to cellular damage. Biochimica et Biophysica Acta - Molecular Cell Research 2009;1793(4):684-696.

Bao M, Booth JL, Elmendorf BJ, Canfield WM. Bovine UDP-N-acetylglucosamine: Lysosomal-enzyme N-Acetylglucosamine-1-phosphotransferase. J Biol Chem 1996b;271(49):31446-51.

Bargal R, Zeigler M, Abu-Libdeh B, Zuri V, Mandel H, Ben Neriah Z, et al. When Mucolipidosis III meets Mucolipidosis II: GNPTA gene mutations in 24 patients. Mol Genet Metab 2006;88(4):359-63.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV. The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution 2003;18(5):249-256.

Cathey SS, Kudo M, Tiede S, Raas-Rothschild A, Braulke T, Beck M, Taylor HA, Canfield WM, Leroy JG, Neufeld EF, McKusick VA. Molecular order in mucolipidosis II and III nomenclature. Am J Med Genet Part A 2008;(146A):512–513.

Cathey SS, Leroy JG, Wood T, Eaves K, Simensen RJ, Kudo M, Friez MJ. Phenotype and genotype in mucolipidoses II and III alpha/beta: a study of 61 probands. Journal of Medical Genetics 2010;47(1):38-48.

Coutinho MF, Alves S. From rare to common and back again: 60 years of lysosomal dysfunction. Mol Genet Metab 2016;117(2):53-65.

Coutinho MF, Encarnação M, Gomes R, da Silva Santos L, Martins S, Sirois-Gagnon D et al. Origin and spread of a common deletion causing mucolipidosis type II: Insights from patterns of haplotypic diversity. Clin Genet 2011;80(3):273-80.

Cury GK, Matte U, Artigalás O, Alegra T, Velho R V., Sperb F et al. Mucolipidosis II and III alpha/beta in brazil: Analysis of the GNPTAB gene. Gene 2013;524(1):59-64.

El Husny AS, Caldato MCF. Erros Inatos Do Metabolismo: Revisão De Literatura. Rev Para Med 2006;20(2):41-5.

Encarnação M, Lacerda L, Costa R, Prata MJ, Coutinho MF, Ribeiro H, et al. Molecular analysis of the GNPTAB and GNPTG genes in 13 patients with mucolipidosis type II or type III - Identification of eight novel mutations. Clin Genet 2009;76(1):76-84.

Filocamo M, Morrone A, Metaboliche PM, Neuroscienze D, Gaslini IG, Gaslini LG. Lysosomal storage disorders: Molecular basis and laboratory testing. Hum Genomics 2011;5(3):156-159.

Franke M, Braulke T, Storch S. Transport of the GlcNAc-1-phosphotransferase α/β-subunit precursor protein to the golgi apparatus requires a combinatorial sorting motif. J Biol Chem 2013;288(2):1238-49.

Futerman AH, Van Meer G. The cell biology of lysosomal storage disorders. Nat Rev Mol Cell Biol 2004;5(7):554-65.

Giugliani R, Federhen A, Michelin-tirelli K, Riegel M. Relative frequency and estimated minimal frequency of Lysosomal Storage Diseases in Brazil: Report from a Reference Laboratory. Genet Mol Biol 2017

Hashemi-Gorji F, Ghafouri-Fard S, Salehpour S, Yassaee VR, Miryounesi M. A novel splice site mutation in the GNPTAB gene in an Iranian patient with mucolipidosis II α/β. J Pediatr Endocrinol Metab 2016;29(8):991-3.

Hers, HG. α-Glucosidase deficiency in generalized glycogen-storage disease (Pompe's disease). Biochemical Journal, 1963;1(86):11-16.

Hu YB, Dammer EB, Ren RJ, Wang G. The endosomal-lysosomal system: From acidification and cargo sorting to neurodegeneration. Transl Neurodegener Translational Neurodegeneration 2015;4(1):1-10.

Huotari J, Helenius A. Endosome maturation. EMBO J Nature Publishing Group 2011;30(17):3481-500.

Kang C, Drayna D. A role for inherited metabolic deficits in persistent developmental stuttering. Molecular Genetics and Metabolism 2010;107(3):276-280.

Kingma SDK, Bodamer OA, Wijburg FA. Epidemiology and diagnosis of lysosomal storage disorders; Challenges of screening. Best Pract Res Clin Endocrinol Metab 2015;29(2):145-57.

Koehne T, Markmann S, Schweizer M, Muschol N, Friedrich RE, Hagel C et al. Mannose 6-phosphate-dependent targeting of lysosomal enzymes is required for normal craniofacial and dental development. Biochim Biophys Acta - Mol Basis Dis 2016;1862(9):1570-80.

Kornfeld S, Mellman I. The Biogenesis of Lysosomes. Annu Rev Cell Biol 1989;5(1):483-525.

Ludwig NF. Análise do gene *GNPTAB* em pacientes brasileiros com Mucolipidose II/III. 92p. Dissertação (Mestrado em Ciências Médicas) – Faculdade de Medicina, Universidade Federal do Grande do Sul, Porto Alegre, 2016.

Luzio JP, Hackmann Y, Dieckmann NMG, Griffiths GM. The biogenesis of lysosomes and lysosome-related organelles. Cold Spring Harb Perspect Biol 2014;14(10):1-18.

Luzio JP, Pryor PR, Bright NA. Lysosomes: Fusion and function. Nat Rev Mol Cell Biol 2007;8(8):622-32.

Meikle P, Hopwood J, Clague A, Carey W. Prevalence of lysosomal storage disorders. Jama 1999;281(3):249-54.

Moammar H, Cheriyan G, Mathew R, Al-Sannaa N. Incidence and patterns of inborn errors of metabolism in the Eastern Province of Saudi Arabia, 1983-2008. Ann Saudi Med 2010;30:271-271.

Plante M, Claveau S, Lepage P, Lavoie ÈM, Brunet S, Roquis C et al. Mucolipidosis II: A single causal mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTAB) in a French-Canadian founder population. Clin Genet 2008;73(3):236-44.

Qian Y, Flanagan-Steet H, van Meel E, Steet R, Kornfeld SA. The DMAP interaction domain of UDP-GlcNAc: lysosomal enzyme N-acetylglucosamine-1-phosphotransferase is a substrate recognition module. Proc Natl Acad Sci 2013;110(25):10246-51.

Rao AN, Kavitha J, Koch M, Kumar S V. Inborn Errors of Metabolism: Review and data from a Tertiary Care Center. Indian J Clin Biochem 2009;24(3):215-22.

Raza MH, Domingues CEF, Webster R, Sainz E, Paris E, Rahn R, et al. Mucolipidosis types II and III and non-syndromic stuttering are associated with different variants in the same genes. Eur J Hum Genet. Nature Publishing Group 2016;24(4):529-34.

Rohrer J, Kornfeld R. Lysosomal hydrolase mannose 6-phosphate uncovering enzyme resides in the trans-Golgi network. Mol Biol Cell 2001;12(June):1623-31.

Sandhoff K, Kolter T. Topology of glycosphingolipid degradation. Trends Cell Biol 1996;6(March):98-103.

Sanseverino MT, Wajner M, Giugliani R. Application of a clinical and laboratory protocol for the investigation of inborn errors of metabolism among critically ill children. J Pediatr (Rio J) 2000;76(5):375-82.

Schwake M, Schröder B, Saftig P. Lysosomal membrane proteins and their central role in physiology. Traffic 2013;14(7):739-48.

Settembre C, Ballabio A. Lysosomal adaptation: How the lysosome responds to external cues. Cold Spring Harb Perspect Biol 2014;6(6):1-15.

Slatkin M. Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. Nat Rev Genet 2008;9(6):477-85.

Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M et al. Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate transferase α- and β-subunit (GNPTAB) gene mutations causing mucolipidosis types IIα/β and IIIα/β in 46 patients. Hum Mutat 2009;30(11):956-73.

Tay, W. Symmetrical changes in the region of the yellow spot in each eye of an infant, Trans. Opthalmol. Soc. 1881;(1):55-57.

Tettamanti G, Bassi R, Viani P, Riboni L. Salvage pathways in glycosphingolipid metabolism. Biochimie 2003;85(3-4):423-37.

Tiede S, Storch S, Lübke T, Henrissat B, Bargal R, Raas-Rothschild A et al. Mucolipidosis II is caused by mutations in GNPTA encoding the α/β GlcNAc-1-phosphotransferase. Nat Med 2005;11(10):1109-12.

Turchetto-zolet AC, Turchetto C, Zanella CM, Passaia, G. Marcadores moleculares na era genômica: metodologias e aplicações. Ribeirão Preto: Sociedade Brasileira de Genética, 2017.181p.

van Meel E, Kornfeld S. Mucolipidosis III GNPTG Missense Mutations Cause Misfolding of the γ Subunit of GlcNAc-1-Phosphotransferase. Hum Mutat 2016;37(7):623-6

4 ARTIGO

Os dados deste estudo serão apresentados em forma de artigo científico a ser submetido à revista *Clinical Genetics Wiley*. Tabelas e figuras são apresentadas no final do artigo.

GNPTAB ANALYSIS: HAPLOTYPE CONSTRUCTION AND ORIGIN OF THE MOST COMMON PATHOGENIC MUTATION CAUSING MUCOLIPIDOSIS II AND III ALPHA/BETA IN THE BRAZILIAN POPULATION

(Short running title: ML II/III α/β in Brazil: Haplotypic analysis)

Malu Bettio Soares[1]
Fernanda Sperb-Ludwig[2,3]
Ida VD Schwartz[1,2,3,4]

ACKNOWLEDGMENTS

This study was supported by the Brazilian Research Agency (CNPq), FAPERGS (Rio Grande do Sul Research Support Foundation) and FIPE-HCPA. Samples were obtained from the Inborn Errors of Metabolism Center in Clinical Hospital of Porto Alegre (EIM-HCPA). Professor Dr. Nelson Fagundes provided extremely helpful information about the software PHASE.

[1] Federal University of Rio Grande do Sul (UFRGS)
[2] Post-Graduation Program in Genetics and Molecular Biology (PPGBM-UFRGS)
[3] Medical Genetics Service, Clinical Hospital of Porto Alegre (HCPA), Brazil
[4] Genetics Department, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil

ABSTRACT

N-acetylglucosamine-1-phosphotrasnferase (GlcNac-1-phosphotransferase) is a hexameric enzyme complex encoded by two genes: *GNPTAB* and *GNPTG*. This enzyme has a key role in lysosomal hydrolase trafficking, which is impaired in Mucolipidosis II/III alpha/beta, caused by biallelic mutations in *GNPTAB,* or III gamma, caused by biallelic mutations in *GNPTG*. c.3503_3504delTC is the most frequent mutation in Mucolipidosis II/III alpha/beta. This study aims to characterize the haplotypes found in Brazilian ML II/III alpha/beta patients presenting the c.3503_3504delTC mutation. DNA extraction from blood samples was performed and 9 regions (6 introns and 3 exons) of the *GNPTAB* gene were PCR-amplified and sequenced according to samples availability. Fifteen patients (2 consanguineous) and 100 controls were analyzed. Patients were from several regions of Brazil and controls were from Porto Alegre (RS). Software analysis allowed a network construction to evaluate relations between haplotypes found: 11 in patients (7 bearing c.3503_3504delTC) and 28 in controls. The most common haplotype in patients was found widely distributed in all regions, and a core haplotype shared by 6/7 c.3503_3504delTC bearing haplotypes was identified. Linkage Disequilibrium analysis showed strong LD between the core haplotype markers. No founder effect was proven in Brazil, but haplotypes containing c.3503_3504delTC seem to have a common origin.

Keywords: Mucolipidosis II, Mucolipidosis III Alpha Beta, Haplotypes, mutation, polymorphism.

INTRODUCTION

Mucolipidosis types II (MLII or I-cell disease; OMIM #252500) and III alpha/beta or III gamma (MLIII alpha/beta and ML III gamma or pseudo-Hurler polydystrophy; OMIM #252600) are lysosomal disorders that occur as a result of defective or absent N-acetylglucosamine-1-phosphotransferase's (E.C.2.7.8.17; referred to as GlcNAc-1-phosphotransferase) activity. ML II and III are rare autosomal recessive diseases, with

ML II presenting estimated prevalence of 1:123,500 live births in Portugal[1], 1:252,500 in Japan[2], 1:625,500 in the Netherlands[3] and 1.56:100,000 in Ireland[4] (Northern Ireland). An unusually high prevalence of ML II was found in a small population in Canada[5].

GlcNAc-1-phosphotransferase has a key role in lysosomal hydrolase trafficking, which is impaired in these conditions, leading to accumulation of non-degraded macromolecules and intercellular deficiency of lysosomal enzymes. Biochemically, these diseases are characterized by the failure to modify lysosomal proteins with Manose-6-phosphate (M6P) residues, which are generated in the Golgi apparatus by the sequential action of two enzymes: GlcNAc-1-phosphotransferase and N-acetylglucosamine- 1-phosphodiester α-N-acetylglucosaminidase (referred to as *Uncovering Enzyme* - E.C. 3.1.4.45).[6-8]

ML II alpha/beta is the most severe form, presenting a fast-progressive course and causing death usually in the first decade of life due to cardiorespiratory failure. It is characterized by its early onset of symptoms that can manifest early after birth or even in prenatal period. Those symptoms include impaired skeletal growth, psychomotor retardation, short stature, coarse facial features, gingival hypertrophy, macroglossia, inguinal hernias, skin infiltrates, restricted joint range of motion and hearing loss.[7-9] ML III alpha/beta and ML III gamma are mildest forms, presenting a slower progression course, later onset of clinical symptoms usually during the first childhood and a higher life expectancy up to the eighth decade of life. Symptoms include progressive joint stiffness, claw hands, carpal and tarsal tunnel syndrome, scoliosis and decreased knees and hip joints mobility.[10,11]

The MLII/III diagnosis relies on a combination of clinical suspicion and confirmatory testing and it is based on clinical, radiological, biochemical and molecular findings. Biochemical analysis can be performed through measurement of plasma lysosomal enzyme levels and detection of deficient intracellular enzyme activity. Patients present high levels of lysosomal hydrolase in plasma and reduced concentration of lysosomal hydrolases inside the cell.[12,13]

GlcNAc-1-phosphotransferase catalyzes the transfer of N-acetilglucosamine-1-phosphate to C6 of specific mannose residues in high mannose-type oligosaccharides

on newly-synthesized lysosomal enzymes[6,7], which is the first step in generating M6P recognition marker required for efficient targeting of most lysosomal hydrolases.[14-16] In the second step, the *Uncovering Enzyme*, encoded by the *NAGPA* gene, removes the N-acetylglucosamine cover, exposing the M6P residue. This modification allows recognition of lysosomal hydrolases by two types of receptors and correct addressing to lysosome.[7]

GlcNAc-1-phosphotrasnferase is a hexameric complex formed by $\alpha_2\beta_2\gamma_2$ encoded by two genes. *GNPTAB* gene is located in position 12q23.3, spans 85kb and is composed by 21 exons. It encodes the precursor protein of subunits α and β of N-acetylglucosamine-1-phosphotransferase, a protein with 1256 amino acids and molecular mass of 144kDa. The γ subunits are encoded by *GNPTG*, which spans 11.13kb in 16p13.3, presenting a molecular mass of 34kDa.[6,17,18]

Pathogenic mutations in the *GNPTAB* gene are associated with ML II/III alpha/beta and mutations in the *GNPTG* gene have been associated with ML III gamma. About 160 (HGMD) mutations have been described for *GNPTAB,* 153 of which cause MLs and 7 which are associated with stuttering. For *GNPTG*, there are 42 mutations of which 26 cause ML III gamma and 15 are associated with stuttering (one is associated with retinitis pigmentosa and spondyloepiphyseal dysplasia). In a study performed by Raza[19], it was shown that mutations causing MLs are fundamentally different from those associated with stuttering.[7,14,20]

Pathogenic mutations associated with MLs are rare and most are private, however, c.3503_3504delTC (rs34002892, p.L1168QfsX5) was reported as the most frequent pathogenic mutation associated with MLII around the world. The high frequency of occurrence of one pathogenic mutation among a group of patients can be attributed to a founder effect or to the existence of a mutational hotspot in that area of the gene.[21] Some properties of intrinsic mutation hotspots such as repetitive DNA sequences (in this case TCTC) are related to mutagenesis processes.[22] c.3503_3504delTC was also associated with a founder effect in a French-Canadian population.[5,23] It is located in exon 19 of the *GNPTAB* gene and has been found in high frequencies among patients diagnosed with MLs II and III alpha/beta in distinct populations.[24]

It was identified in 51% in Italian patients[15]; 22% in the United States population[10], in 45% of Portuguese patients[6] and Bargal[9] found 50% alleles with the deletion of patients analyzed from an Arab-Muslim origin. It was also the only pathogenic mutation identified in a sample from Saguenay-Lac-St-Jean population from Canada.[5] In Brazil, it has been found an allelic frequency of 45% in 12 patients[20]. Ludwig[25] found an allelic frequency of 40%, similar to the findings by Cury.[20]

In this scenario, we characterize the haplotypes of 15 patients with ML II and ML III alpha/beta presenting c.3503_3504delTC and investigate the origin of this pathogenic mutation in Brazilian population.

MATERIALS AND METHODS

Fifteen Brazilian ML II/III patients were analyzed (two with consanguineous parents), all compound heterozygous or homozygous for c.3503_3504delTC[20,25]. Alleles from patients with consanguineous parents were considered as one. A sample of 100 healthy anonymous blood donors from Porto Alegre (South Brazil) was also analyzed. All 100 samples were screened for c.3503_3504delTC and do not presented it.

Patients were from different locations in Brazil: three from Bahia (BA), two from Ceará (CE), two from Piauí (PI), two from Rio de Janeiro (RJ), two from Mato Grosso (MT), one from Alagoas (AL), one from Santa Catarina (SC) and one from Rio Grande do Sul (RS).

In addition to pathogenic mutation c.3503_3504delTC, nine *GNPTAB* markers with MAF between 31%-90% (table 1) were selected to compose the haplotype analysis. In controls, only c.365+96_365+97delGT (rs4015837) and c.365+145C>T (rs2108694) in intron 4, c.1285-166G>A (rs7963747) in intron 10, c.1932A>G (rs10778148) in exon 13, c.3135+5T>C (rs759935) in intron 15 and c.3336-25T>C (rs3736476) in intron 17 were analyzed.

Polymerase Chain Reaction (PCR) amplified target samples and was performed as described by Cury.[20] PCR products were purified with a Polyethylene Glycol (PEG) 8000 NaCl 20% solution and separated in an ABI PRISM 3500 Genetic Analyzer.

Sequencing results were analyzed using Chromas software and compared with the main NCBI database sequence for *GNPTAB* gene NG_021243.1 (RefSeqGene).

In homozygous patients for c.3503_3504delTC and other markers, haplotype phase was directly inferred. In heterozygous patients, the phase was determined by parental screening (20 parents had distinct exons and introns analyzed – all parents of heterozygous patients for c.3503_3504delTC had the exon 19 sequenced) when possible due to sample availability and by PHASE (V2.1.1)[26] software.

PHASE run with 100x burn-in and 100 thinning, run multiple times and results were compared in order to analyze reliability (results' stability). To assist in heterozygous phasing, an input file was made specifying heterozygous and homozygous (known and unknown phases) for each position, so haplotypes were inferred correctly and with more reliability.

MEGA software (V 6.0)[27] was used for pairwise and multiple alignments using the implemented function ClustalW.[28] Sequences were adjusted after alignment to maintain only one gap in each deletion in order to avoid incorrect data.

DnaSP[29] was used for the analysis of DNA polymorphisms using data from several *loci*. The haplotypes were constructed in DnaSP software considering variation at nine SNP markers: c.-41_-39delGGC, c.18G>A, c.27G>A, c.365+96_97delTC, c.365+145C>T, c.1285-166G>A, c.1932A>G, c.3135+5T>C and c.3336-25T>C. These polymorphisms are referred to in dbSNP[34] as previously cited, however, c.365+145C>T, c.1285-166G>A, c.1932A>G and c.3135+5T>C presented the ancestral alleles as the opposite of their representation in dbSNP. Arlequin (V 3.5.2.2)[30] was used to calculate genetic diversity, linkage disequilibrium (with two markers from Exon 1 excluded automatically by the software) and performed neutrality tests (Tajima's D- which evaluates the neutral mutation hypothesis, but may depend of sample size, and Fu's FS - which also evaluates the neutral mutation hypothesis, but is more independent of sample size)[31,32] intra-groups and inter-groups (total). The groups were patient's region of residence: Northeast, Midwest, Southeast and South of Brazil.

The software NETWORK was used to form haplotype networks in alleles presenting the deletion by *Median-joining network*[34], implemented in the software.

Networks constructed for patients and controls were in separate files due to the difference of markers analyzed.

Statistical analysis of allelic frequencies of controls, patients and a database frequency[35] was made using the WINPEPI software for Windows (p value>0,05 was considered significant).

RESULTS

Inferred reconstructed haplotypes from the 15 patients (28 alleles) are presented on table 2.

Seven haplotypes were found in 18 chromosomes presenting c.3503_3504delTC from a total of 11 haplotypes. This deletion appears in 18 of 28 alleles (64.2%). Haplotype diversity considering all patients was 0.839 (+-0.0484).

One haplotype was predominant: H4 (7:G:G:delGT:T:A:G:C:C:delTC), which was found in 9 alleles from 8 patients, that represent 50% of alleles with c.3503_3504delTC, and 32.1% of total alleles. The second most frequent haplotype (H3) did not present c.3503_3504delTC mutation and was found in 7 alleles from 7 patients (23.3% of total alleles).

From the 7 haplotypes presenting c.3503_3504delTC (18 alleles), 6 (16 alleles) among these c.3503_3504delTC-harbouring haplotypes present a common core of 7:G:G:delGT:T. The other haplotype presents 8:G:G:delGT:T, varying only in c.-41_-39delGGC site. Outside the core, variation occurs in these haplotypes in the last four sites: c.1286-166G>A, c.1932A>G, c.3135+5T>C and c.3336-25T>C. Marker c.365+96_97delGT do not presented variation on patient's alleles harbouring c.3503_3504delTC, but is polymorphic if all patients are considered and in general population.

The haplotypes are presented per region of patient's origin in table 3. Northeast is the region with more different haplotypes (9 haplotypes in 16 alleles), in which six present the pathogenic deletion c.3503_3504delTC. Bahia (BA) has the highest number of patients in Brazil (3/15) and 4 different haplotypes from which 2

have c.3503_3504delTC. Piauí (PI) has 4 different haplotypes from which 3 have c.3503_3504delTC in 2/15 patients. There is no region showing only c.3503_3504delTC haplotypes. Haplotypic diversity for each region is presented in table 3, as well as average nucleotide diversity (π).

Tajima's D for patients was 2.32 (p = 0.99), indicating that these alleles are not under selection. However, analysis within groups show Tajima's D of 2.04 (p = 0.98) for Northwest, -0.21 (p =0.55) for Midwest, -0.14 (p = 0.45) for Southeast and 0.37 (p = 0.74) for South, suggesting those last three groups are suffering negative selection (table 3)[31]. Results of Fu's FS in all haplotype analysis point to a negative selection, as well as in Northeast and South groups (-1.4 p=0.28, -0.6 p=0.34 and -0.6 p=0.16, respectively). These results are passive to error due to small sample size.

The haplotypes distribution is presented in a map in figure 2. The most frequent haplotype, H4 was widely distributed in all regions. H3, the second most frequent haplotype which does not present c.3503_3504delTC is also well distributed all over Brazil. The results were used to create a network connecting all haplotypes found in patients (figure 2A) and in controls (figure 2B). Those networks are separated due to the difference in number of markers analyzed in both groups.

Network presenting patients' haplotypes shows a close relation of haplotypes carrying the pathogenic deletion. H4 is the central haplotype, distant only one mutational step away from H6, H7 and H10, and two steps from H1, from which only H6 do not present c.3503_3504delTC. This network illustrated more distant relations (more mutational events of distance) of H4 with haplotypes who do not present c.3503_3504delTC (H5, H11 and H3), being closer only to H6. Linkage disequilibrium (LD) was calculated in Arlequin software between all loci within each other. The *loci* presented LD, except c.-39_-41delGGC with c.3135+5T>C and c.3135+5T>C with c.3336-25T>C (table 4) in tests with all groups. All *loci* presented LD with c.3503_3504delTC except c.3135+5T>C.

Recombination rate (R) resulting in a minimum number of recombination events of three (Rm = 3). Recombination has been detected between the last three polymorphic sites (except c.3503_3504delTC).

In control analysis, only 6 of the 9 markers were analyzed: c.365+96_97delGT, c.365+145C>T, c.1285-166G>A, c.1932A>G, c.3135+5 T>C and c.3336-25 T>C. A total of 28 haplotypes were found, presenting a diversity of 0.92 (p=0.0069).

The most frequent haplotype was HC3 (GT:C:G:A:T:T), present in 27 of 200 alleles (13,5%). The second most frequent haplotype in controls was HC2, similar to H6 found in patients (which does not present the pathogenic deletion – and only differs by the presence of the deletion from the most frequent in patients H4), appearing in a frequency of 13% (26 of 200 alleles). H4 and H6 are both alleles from just one patient, heterozygous for the pathogenic deletion. All haplotypes found and their frequency in control samples are presented in table 5. Control samples showed 4 of 6 alleles with LD (table 6). A network was constructed also for control samples and is presented in figure 3.

Allele frequency at polymorphic sites were compared between patient samples, control samples and database information (from phase 3 of 1000 Genomes Project). To improve analysis of LD, a statistical analysis comparing the three sets are presented in figure 4 Mutated alleles appeared in higher frequencies in patients except c.1285-166G>A and c.3135+5T>C, but only two had significant difference when compared with controls (c.365+96_97delGT and c.365+145C>T, both located at intron 4). In comparison with 1000 Genomes Project database, only c.365+96_97delGT and c.3336-35T>C had significant difference.

DISCUSSION

Several studies performing haplotype analysis are being published and there is a growing interest in understanding haplotype structures in the human genome using genetic markers. Due to the development of molecular biology techniques, analysis of hundreds of samples in populations is now possible. Haplotype structures can provide information on human evolutionary history and identification of genetic variants underlying various traits. One of the main objectives of haplotype analysis is to identify

Linkage Disequilibrium, the non-random allele association that provides information to infer populational history.[37]

In this study, haplotype construction and LD analysis of 15 Brazilian patients with ML II/III alpha/beta and 100 control samples was performed.

Haplotype construction was made considering nine polymorphic sites, three of which have been used for haplotype construction and analysis of the same common pathogenic mutation (c.3503_3504delTC), as in the work of Coutinho[23]: c.-41_-39delGGC, c.18G>A and c.1932A>G. Their group found 9 haplotypes in 44 patients from six different countries, including 3 of our patients, different from our findings of 11 haplotypes only from Brazilian population. The haplotype diversity values are higher in Brazilian population: 0.839 and in their analysis, the obtained value was 0.709, that is not significant.

The most frequent haplotype found in Brazilian ML patients was H4, the same haplotype found by Coutinho[23] when considering common markers (7:G:G), their work analyzed 3 Brazilian patients. In Coutinho[23]'s work 70% alleles presented this haplotype and, considered as a core haplotype, it was found in 97% of all chromosomes analyzed by them. In their work, the core haplotype 7:G:G was shared by Italian, Portuguese and Israeli Arab-Muslim patients, peaking the high frequency of 70% in Italians and 37.5% in Portuguese. This core haplotype is presented in all regions of Brazil in at least one allele. It is the same partial haplotype present in Coutinho[23]'s work. Rio Grande do Sul in South region (RS: n = 1) had only H4 presenting the deletion and H6, which is similar except for c.3503_3504delTC. This can explain probably due to the strong European and especially Italian heritage that occurs in Rio Grande do Sul. According to Moura[38], South region has a composition of 77% of European ancestry, 12% of African and 11% American, the biggest European proportional composition in Brazil.

The most frequent haplotype in controls, HC3, on which all variant polymorphisms are absent, is precisely distinct from the most frequent haplotype found in chromosomes presenting c.3503_3504delTC, which could indicate that there is indeed a relation between those positions and the deletion in Brazilian population, which is partially corroborated by LD analysis. On the other hand, CH2, the second

more frequent haplotype in controls, is composed by the last six polymorphisms present in H4 and similar to H6 except for the mutation c.3503_3504delTC, which could indicate that this specific haplotype could have acquired the pathogenic mutation. Baiotto[39] suggests that haplotypes found in lower frequency in relation to one predominant haplotype indicate that they arose from separated events and suffered recombination.

The high number of haplotypes found in controls match the number of haplotypes described in controls for the core haplotypes, comparing common markers in literature. Murdoch[40] had also reduced main haplotypes to core haplotypes to further analysis of historical background. Coutinho[23] found 35 haplotypes in control samples of Portuguese population with the same core haplotype (only G:G). The highest frequency found in our controls was of 13.5%, matching the 14% found by Coutinho.[23] Their core haplotype was found in controls in 27%, contrasting with their high percentages in patients, with 46% presenting full haplotype and 97% presenting core haplotype. In our study, the core haplotype 7:G:G:delGT:T found in patients could only be compared with control samples if reduced to delGT:T due to absence of the first 3 positions in control analysis. This partial haplotype was present in only six haplotypes of 28 in controls (21%), contrasting with the 81% (reduced core haplotype) and 72% (extended core haplotype) found in patients. These findings are compatible with the results of Coutinho.[23]

H4 was distributed in patients all over Brazil and the other haplotypes presenting the pathogenic deletion show a core haplotype, with little variation at some markers. According to Lerner-Ellis[41], this configuration is consistent with a single-occurring background event, followed by mutation or recombination modifying the ancestral haplotype. Coutinho[23] sustain the hypothesis of a single-occurring event and strong linkage between the three markers analyzed by them and c.3503_3504delTC.

Results of Tajima's D and Fu's FS suggest for regions Northeast, Midwest and Southeast that a negative selection is occurring. In the other regions, Tajima's D and Fu's FS positive (not significant) value demonstrates that no selection is occurring.[42] It is expected for pathogenic mutations such as c.3503_3504delTC, which cause a severe phenotype, that negative selection pressure and purifying selection occur, but

due to reduced sample size, this result can be masked to cause false positives/negatives. On the other hand, the different haplotypes bearing the pathogenic mutation seem to be arising from distinct events in Brazilian population.

The networks show the same relation with haplotypes bearing c.3503_3504delTC and haplotypes similar to them in control samples. In opposite to Coutinho[23]'s work, haplotypes show a close relation within chromosomes bearing c.3053_3504delTC and a more distant relation with chromosomes who do not. It is noted a reticular pattern in patient's network in chromosomes with the pathogenic deletion. According to Arenas[43], a reticulated pattern is quite common in population studies, usually accommodating events like recombination, hybridization or lateral gene transfer. So, it would be possible to see the haplotypes that appeared by recombination events from H4 and appear in this reticular pattern (H7, H8, H9, H10).

Coutinho[23] attributed a single-occurring event and indicated that c.3503_3504delTC arose from a single population through founder effect in a peri-Mediterranean region of Europe and then have spread throughout the continent due to migration. A widespread unique haplotype and haplotypes closely related corroborate this information. According to their study, the massive emigration of Europeans to the New World is responsible for the mutation transport to the continent. Around 500.000 Portuguese have arrived in Brazil from 1500 to 1808 and in the period from 1872 to 1975, Brazil received at least 5.5 million other immigrants from Europe and other parts of the world: 34% Italians, 29% Portuguese, 14% Spanish, 5% Japanese, 4% Germans, 2% Lebanese and Syrians, and 12% others[45]. Coutinho[23] attest that North and South American haplotypes fall into the category of the European founder haplotype leading to the idea that probably, in Brazil, this mutation is also a product of the same single event, which arose in Europe. Due to this massive immigration and distinct proportional ancestry indices in Brazil, we can infer that recombination events may have constructed more diverse haplotypes bearing c.3503_3504delTC in this highly admixed population.[38]

There was no founder effect detected in Brazil or in the Brazilian population for this mutation, probably due to several events of migration reported, and the haplotypes maintain the same relations in control and patient's samples. However, we cannot

discard the possibility of one common event occurring as the origin of this mutation in Brazilian population, since our results are similar as the ones found in Coutinho's work and the ancient age attributed to the mutation tend to lead us to the same conclusions as theirs.

A population's demographic history leaves a genetic mark that can be seen throughout its evolutionary history and to estimate haplotype frequencies in each group separately to see if whether trait-influencing variants reside within or near the genomic region spanned by the markers is a reasonable and reliable method of evaluating the course and evolution of pathogenic mutations. By determining the extended haplotypes at any given locus in a population, we can identify which SNPs will be redundant and which will be essential in association studies.[45]

This study determined haplotypes of patients presenting c.3503_3504delTC, a pathogenic mutation of extreme relevance in this population, once it's responsible for the majority of ML II/III alpha/beta cases in Brazil.[20,25,46] This analysis enables an omnibus comprehension of mechanisms that originated this mutation in Brazilian population, besides being an important tool of molecular diagnose for those diseases.

REFERENCES

1. Pinto R, Caseiro C, Lemos M, Lopes L, Fontes A, Ribeiro H, Pinto E, Silva E, Rocha S, Marcao A, Ribeiro I, Lacerda L, Ribeiro G, Amaral O, Sa Miranda MC. Prevalence of lysosomal storage diseases in Portugal. Eur J Hum Genet 2004;12:87-92.

2. Okada S, Owada M, Sakiyama T, Yutaka T, Ogawa M. I-cell disease: clinical studies of 21 Japanese cases. Clin Genet 1985;28:207-15.

3. Poorthuis BJHM, Wevers R, Kleijer WJ, Groener JEM, de Jong JGN, van Weely S, Niezen-Koning KE, Van Diggelen OP. The frequency of lysosomal diseases in The Netherlands. Hum Genet 1999;105:151-6.

4. McElligott F, Beatty E, O'Sullivan S, Hughes J, Lambert D, Cooper A, Crushell E. Incidence of I-cell disease (mucolipidosis type II) in the irish population. J Inherit Metab Dis 2011;34(3):S206.

5. Plante M, Claveau S, Lepage P, Lavoie ÈM, Brunet S, Roquis C, Laprise C. Mucolipidosis II: a single causal mutation in the N-acetylglucosamine-1-phosphotransferase gene (GNPTAB) in a French-Canadian founder population. Clinical Genetics 2008;73(3):236-244.

6. Encarnação M, Lacerda L, Costa R, Prata MJ, Coutinho MF, Ribeiro H, Alves S. Molecular analysis of the GNPTAB and GNPTG genes in 13 patients with mucolipidosis type II or type III - Identification of eight novel mutations. Clinical Genetics 2009;76(1):76-84.

7. De Pace R, Coutinho MF, Koch-Nolte F, Haag F, Prata MJ, Alves S, Pohl S. Mucolipidosis II-related mutations inhibit the exit from the endoplasmic reticulum and proteolytic cleavage of GlcNAc-1-phosphotransferase precursor protein (GNPTAB). Human Mutation 2014;35(3):368-376.

8. Koehne T, Markmann S, Schweizer M, Muschol N, Friedrich RE, Hagel C, Braulke T. Mannose 6-phosphate-dependent targeting of lysosomal enzymes is required for normal craniofacial and dental development. Biochimica et Biophysica Acta - Molecular Basis of Disease 2016;1862(9);1570-1580.

9. Bargal R, Zeigler M, Abu-Libdeh B, Zuri V, Mandel H, Ben Neriah Z, Raas-Rothschild A. When Mucolipidosis III meets Mucolipidosis II: GNPTA gene mutations in 24 patients. Molecular Genetics and Metabolism 2006;88(4):359-363.

10. Cathey SS, Leroy JG, Wood T, Eaves K, Simensen RJ, Kudo M, Friez MJ. Phenotype and genotype in mucolipidoses II and III alpha/beta: a study of 61 probands. Journal of Medical Genetics 2010;47(1):38-48.

11. Tüysüz B, Kasapçopur Ö, Alkaya DU, Şahin S, Sözeri B, Yeşil G. Mucolipidosis type III gamma: Three novel mutation and genotype-phenotype study in eleven patients. Gene 2018;642:398-407.

12. Coutinho MF, Encarnação M, Laranjeira F, Lacerda L, Prata MJ, Alves S. Solving a case of allelic dropout in the GNPTAB gene: implications in the molecular diagnosis

of mucolipidosis type III alpha/beta. Journal of Pediatric Endocrinology and Metabolism 2016;29(10):1225-1228.

13. Yang M, Cho SY, Park H-D, Choi R, Kim Y-E, Kim J, Jin D-K. Clinical, biochemical and molecular characterization of Korean patients with mucolipidosis II/III and successful prenatal diagnosis. Orphanet Journal of Rare Diseases 2017;12(1):11.

14. Hashemi-Gorji, F, Ghafouri-Fard S, Salehpour S, Yassaee VR, Miryounesi M. A novel splice site mutation in the GNPTAB gene in an Iranian patient with mucolipidosis II α/β. Journal of Pediatric Endocrinology and Metabolism 2016;29(8):991-993.

15. Tappino B, Chuzhanova NA, Regis S, Dardis A, Corsolini F, Stroppiano M, et al. Molecular characterization of 22 novel UDP-N-acetylglucosamine-1-phosphate transferase α- and β-subunit (GNPTAB) gene mutations causing mucolipidosis types II α/β and III α/β in 46 patients. Hum Mutat 2009;30(11):956-73.

16. Braulke T, Pohl S, Storch S. Molecular analysis of the GlcNac-1-phosphotransferase. Journal of Inherited Metabolic Disease 2008;31(2):253-257.

17. Bao M, Booth JL, Elmendorf BJ, Canfield WM. Bovine UDP-N-acetylglucosamine: lysosomal-enzyme N-Acetylglucosamine-1-phosphotransferase. J Biol Chem 1996;271(49): 31437-31445.

18. Tiede S, Storch S, Lübke T, Henrissat B, Bargal R, Raas-Rothschild A, Braulke T. Mucolipidosis II is caused by mutations in GNPTA encoding the alpha/beta GlcNAc-1-phosphotransferase. Nature Medicine 2005;11(10):1109-1112.

19. Raza MH, Domingues CEF, Webster R, Sainz E, Paris E, Rahn R, Drayna D. Mucolipidosis types II and III and non-syndromic stuttering are associated with different variants in the same genes. European Journal of Human Genetics 2016;24(4):529-34.

20. Cury GK, Matte U, Artigals O, Alegra T, Velho RV, Sperb F, Schwart IVD. Mucolipidosis II and III alpha/beta in brazil: analysis of the GNPTAB gene. Gene 2013;524(1): 59-64.

21. Liu S, Zhang W, Shi H, Yao F, Wei M, Qiu Z. Mutation analysis of 16 Mucolipidosis II and III alpha/beta Chinese children revealed genotype-phenotype correlations. PLoS One. 2016;11(9):1-12.

22. Nassiri I, Azadian E, Masoudi-Nejad A. A Sequence Motif Associated with Intrinsic Mutation Hot-Spots in Human Cancers. J Proteomics Bioinform 2013;(6):183-186.

23. Coutinho M., Encarnação M, Gomes R, Silva Santos L da, Martins S, Sirois-Gagnon D, Alves S. Origin and spread of a common deletion causing mucolipidosis type II: insights from patterns of haplotypic diversity. Clinical Genetics 2011;80(3):273-280.

24. National Center for Biotechnology Information: dbSNP – the single nucleotide polymorphism database. https://www.ncbi.nlm.nih.gov/projects/SNP/snpref.cgi?rs=34002892 Accessed March 3, 2017.

25. Ludwig NF, Sperb-Ludwig F, Velho RV, Schwartz IVD. Next-generation sequencing corroborates a probable de novo GNPTG variation previously detected by Sanger sequencing. Molecular Genetics and Metabolism Reports 2017;(11):92-93.

26. Stephens, M., Smith, N., and Donnelly, P. American Journal of Human Genetics 2001;68:978-989.

27. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013;30(12):2725-9.

28.Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22(22):4673-80.

29. Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 2009;25(11):1451-2.

30. Excoffier, L. and H.E. L. Lischer. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 2010;(10): 564-567.

31. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123(3):585-95.

32. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics 1993;133(3):693-709.

33. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31(21):3555-3557.

34. Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999 Jan;16(1):37-48.

35. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A et al. A global reference for human genetic variation. Nature 2015;526(7571):68-74.

36. National Center for Biotechnology Information: dbSNP – the single nucleotide polymorphism database. https://www.ncbi.nlm.nih.gov/projects/SNP Accessed December 1, 2017.

37. Zhao H, Pfeiffer R, Gail MH. Haplotype analysis in population genetics and association studies. Pharmacogenomics 2003;4(2):171-8.

38. Moura RR, Coelho AVC, Balbino VQ, Crovella S, Brandão LAC. Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. Am J Hum Biol 2015;27(5):674-80.

39. Baiotto C, Sperb F, Matte U, da Silva CD, Sano R, Coelho JC et al. Population analysis of the GLB1 gene in South Brazil. Genet Mol Biol 2011;34(1):45-8.

40. Murdoch JD, Speed WC, Pakstis AJ, Heffelfinger CE, Kidd KK. Worldwide population variation and haplotype analysis at the serotonin transporter gene SLC6A4 and implications for association studies. Biol Psychiatry 2013;74(12):879-89.

41. Lerner-Ellis JP, Tirone JC, Pawelek PD, Doré C, Atkinson JL, Watkins D et al. Identification of the gene responsible for methylmalonic aciduria and homocystinuria, cblC type. Nat Genet 2006;38(1):93-100.

42. Jensen JM, Villesen P, Friborg RM, Mailund T, Besenbacher S, Schierup MH. Assembly and analysis of 100 full MHC haplotypes from the Danish population. Genome Res 2017;1-11.

43. Arenas M, Valiente G, Posada D. Characterization of reticulate networks based on the coalescent with recombination. Mol Biol Evol 2008;25(12):2517-20.

44. Pena SDJ, Bastos-Rodrigues L, Pimenta JR, Bydlowski SP. DNA tests probe the genomic ancestry of Brazilians. Brazilian J Med Biol Res 2009;42(10):870-6.

45. Johnson GC, Esposito L, Barratt BJ, Smith a N, Heward J, Di Genova G et al. Haplotype tagging for the identification of common disease genes. Nat Genet 2001;29:233-7.

Table 1: Markers selected to compose haplotype analysis, their location in *GNPTAB* gene, populational frequencies in 1000 genomes Project and RS number

Variant	Location	RS number	Allelic frequency (1000 genomes Project n=)
c.-41_-39delGGC	5'UTR	76300806	delGGC=70.2%
c.18G>A	Exon 1	4764655	A=90.2%
c.27G>A	Exon 1	222504	A=90.2%
c.365+96_97delGT	Intron 4	4015837	delGT=33%
c.365+145C>T	Intron 4	2108694	T=68%
c.1285-166G>A	Intron 10	7963747	A=65%
c.1932A>G	Exon 13	10778148	G=65%
c.3135+5T>C	Intron 15	759935	C=69%
c.3336-25T>C	Intron 17	3736476	C=31%

Table 2: Haplotypes found in 15 patients presenting c.3503_3504delTC, number of patients presenting the haplotypes and number of alleles presenting the haplotypes

Haplotype nº	c.-41_-39delGGC	c.18G>A	c.27G>A	c.365+96_97delGT	c.365+145C>T	c.1285-166G>A	c.1932A>G	c.3135+5T>C	c.3336-25T>C	c.3503_3504delTC	Patients presenting haplotype	Alleles presenting haplotype
H1	8	G	G	delGT	T	G	G	C	C	delTC	1	2
H2	7	G	G	delGT	T	G	G	T	C	delTC	2	2
H3	8	G	G	GT	C	G	A	T	T	TC	7	7
H4**	7	G	G	delGT	T	A	G	C	C	delTC	8	9
H5	7	G	G	delGT	T	A	G	C	T	TC	1	1
H6	7	G	G	delGT	T	A	G	C	C	TC	1	1
H7	7	G	G	delGT	T	A	G	T	C	delTC	2	2
H8	7	G	G	delGT	T	G	G	T	T	delTC	1	1
H9	7	G	G	delGT	T	A	G	T	T	delTC	1	1
H10**	7	G	G	delGT	T	A	A	C	C	delTC	1	1
H11	8	G	G	GT	C	G	G	C	T	TC	1	1

† In position c.-41_-39delGGC, 7 represent the presence of deletion and 8 represents its absence

* * Haplotypes marked with two asterisk include patients with consanguineous parents

Table 3 - Haplotypes per region, neutrality tests Tajima's D and Fu's FS. π indicates nucleotide diversity

Region	Nº of haplotypes	Nº of individuals	Haplotypes	Haplotype diversity	p	π	Tajima's D	p	Fu's FS	p
Northeast	9	8	**H1, H2**, H3, **H4, H7**, **H8, H9**, H11, H12	0.85	0.06	3.8	2.04	0.98	-0.6	0.34
Midwest	3	2	H3 **H4, H7**	0.83	0.22	4.16	-0.21	0.55	1.22	0.70
Southeast	4	3	**H3, H4, H5, H10**	0.86	0.13	3.33	-0.14	0.45	0.42	0.56
South	4	2	**H2**, H3, **H4**, H6	1	0.17	4.5	0.37	0.74	-0.6	0.16
All	11	15	all haplotypes	0.85	0.06	3.65	2.32	0.99	-1.4	0.28
Controls	28	100	28 haplotypes	0.92	0.006*	2.9	3.51	0.99	-11	0.004*

† Haplotypes presenting c.3503_3504delTc are in **bold.**

*statiscally significant p value

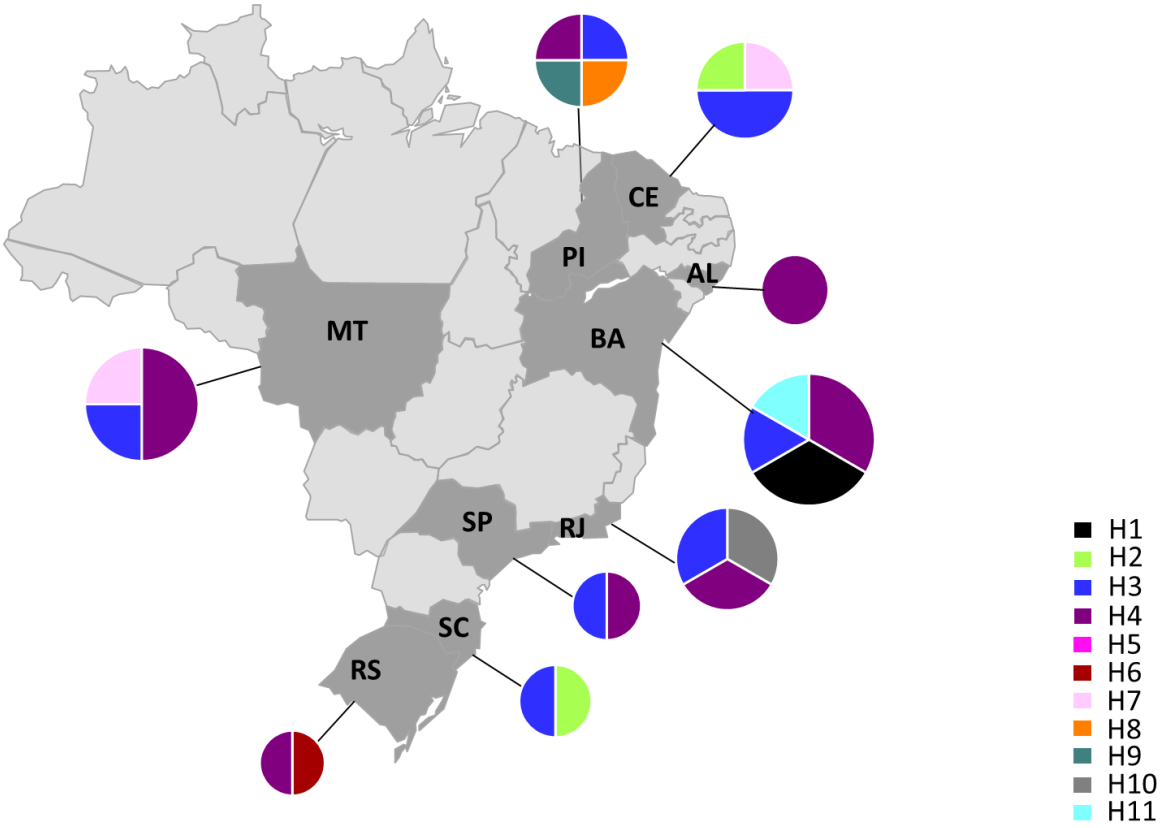

Figure 1. Haplotype distribution in Brazil. The circle size is proportional to the number of patients, which vary from 1 to 3 per region. Colors represent different haplotypes

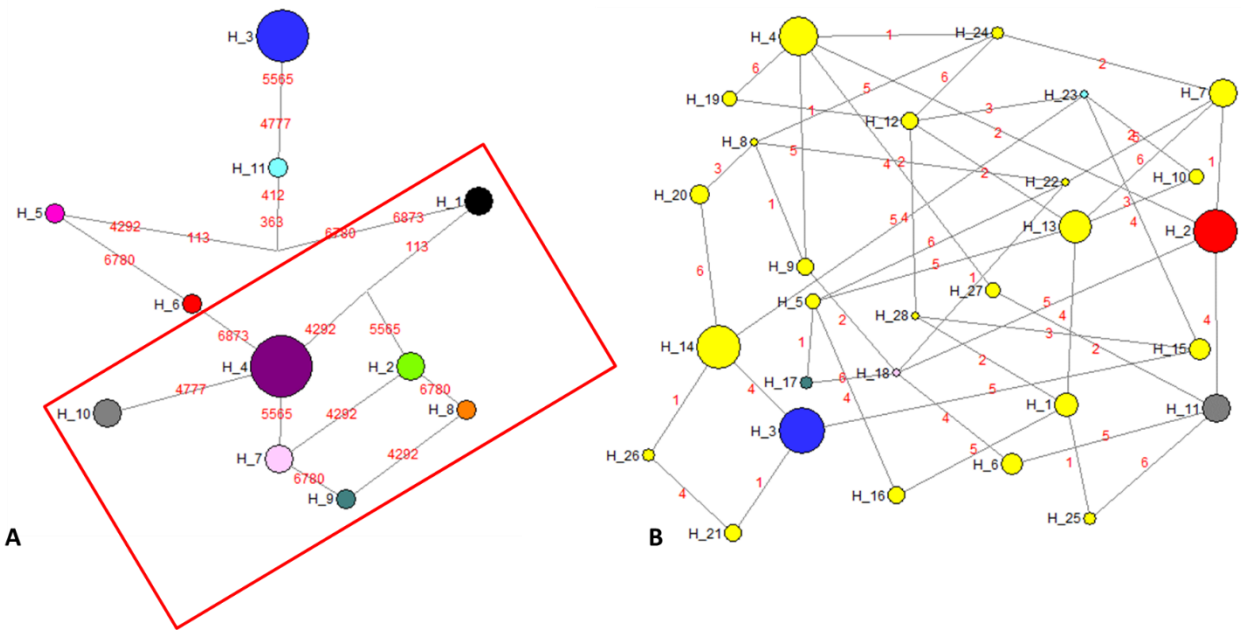

Figure 2: Median-Joining network between eleven haplotypes found in patients (A). Area highlighted inside the red square shows haplotypes presenting c.3503_3504delTC. On the right, a network showing relations between haplotypes in controls (B). Numbers in red between haplotypes are positions of events occurred. Haplotypes match colors with Figure 1

Table 4. Results of Linkage Disequilibrium between all *loci* in the Arlequin software.

Locus	c.-39_-41delGGC	c.365+96_97delGT	c.365+145 C>T	c.1285-166G>A	c.1932A>G	c.3135+5 T>C	c.3336-25T>C	c.3503_3504 delTC
c.-39_-41delGGC	*	+	+	+	+	-	+	+
c.365+96_97delGT	+	*	+	+	+	+	+	+
c.365+145 C>T	+	+	*	+	+	+	+	+
c.1285-166G>A	+	+	+	*	+	+	+	+
c.1932A>G	+	+	+	+	*	+	+	+
c.3135+5 T>C	-	+	+	+	+	*	+	-
c.3336-25T>C	+	+	+	+	+	+	*	+
c.3503_ 3504delTC	+	+	+	+	+	-	+	*

† Loci presenting LD are represented by +, - are *loci* without LD and * are combinations of the same *loci*
p<0,05

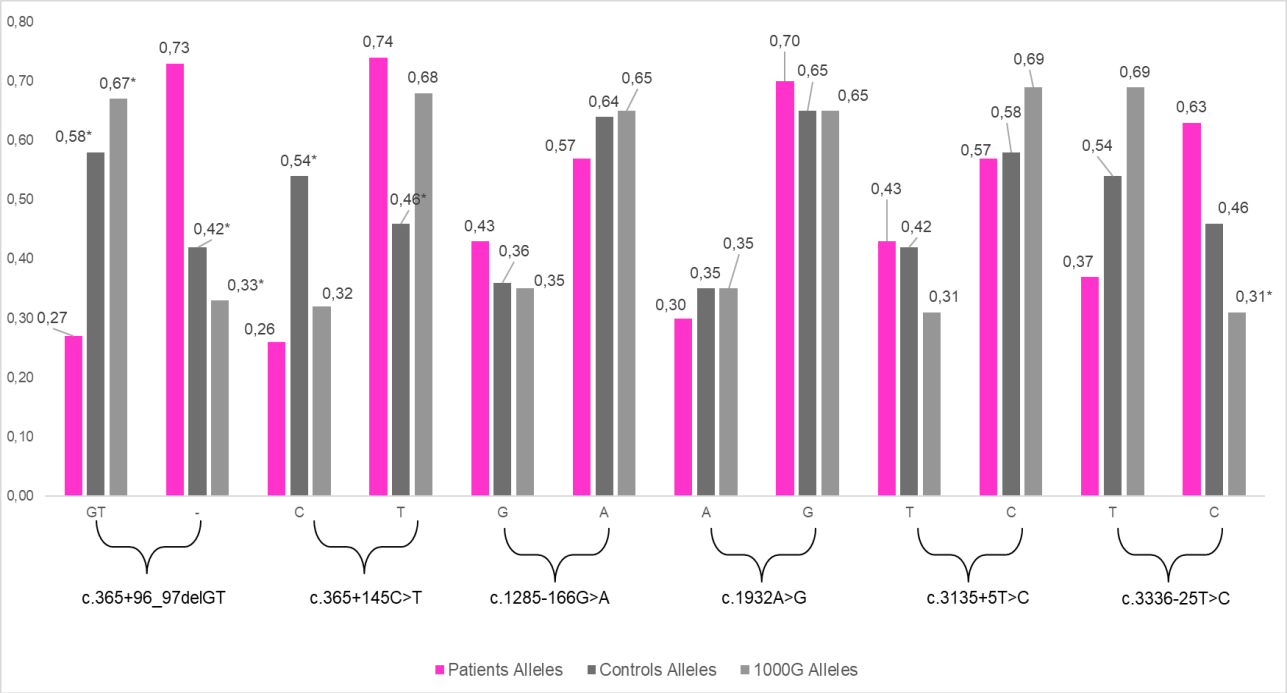

Figure 4: Allelic frequencies at polymorphic sites, comparing patient frequencies, control and database-extracted frequencies (global means, 1000genomes project - phase 3). Each six columns (two sets of one column for patients, one for controls, one for database) are one polymorphic site. Statistical differences were made patients x controls and patients x 1000G and are assigned with an asterisk. Mutated alleles come in second in each set of data

Table 5: Haplotypes found in controls, their frequency and matching patient's haplotypes

Haplotype	c.365+96_97delGT Íntron 4	c.365+145C>T Íntron 4	c.1285-166G>A Íntron 10	c.1932A>G Éxon 13	c.3135+5 T>C Íntron 15	c.3336-25 T>C Íntron 17	Frequency	Common with patients haplotypes
HC1	GT	T	A	A	C	T	4,0%	
HC2	-	T	A	G	C	C	13,0%	H6
HC3	GT	C	G	A	T	T	13,5%	H3
HC4	-	C	A	G	C	C	10,5%	
HC5	GT	T	A	G	T	T	1,5%	
HC6	-	T	A	A	T	C	3,0%	
HC7	GT	T	A	G	C	C	5,5%	
HC8	GT	C	A	G	T	C	0,5%	
HC9	-	C	A	G	T	C	2,0%	
HC10	GT	T	G	G	C	T	1,5%	
HC11	-	T	A	A	C	C	5,0%	H10
HC12	GT	C	A	G	C	T	2,0%	
HC13	GT	T	A	G	C	T	7,5%	
HC14	GT	C	G	G	T	T	12,0%	
HC15	GT	C	G	A	C	T	3,0%	
HC16	GT	T	A	A	T	T	2,0%	
HC17	-	T	A	G	T	T	1,0%	H9
HC18	-	T	A	G	T	C	0,5%	H7
HC19	-	C	A	G	C	T	1,5%	
HC20	GT	C	G	G	T	C	2,5%	
HC21	-	C	G	A	T	T	2,0%	
HC22	GT	T	A	G	T	C	0,5%	
HC23	GT	C	G	G	C	T	0,5%	H11
HC24	GT	C	A	G	C	C	1,0%	
HC25	-	T	A	A	C	T	1,0%	
HC26	-	C	G	G	T	T	1,0%	
HC27	-	C	A	A	C	C	1,5%	
HC28	GT	C	A	A	C	T	0,5%	

Table 6: LD results for control samples for each *locus*

Locus	c.365+96_97delGT	c.365+145C>T	c.1285-166G>A	c.1932A>G	c.3135+5T>C	c.3336-25T>C
c.365+96_97delGT	*	+	+	-	+	+
c.365+145 C>T	+	*	+	-	+	+
c.1285-166G>A	+	+	*	+	+	+
c.1932A>G	-	-	+	*	+	+
c.3135+5 T>C	+	+	+	+	*	+
c.3336-25T>C	+	+	+	+	+	*

† *Loci* presenting LD are represented by +, - are *loci* without LD and * are combinations of the same *loci*
p<0,05