THÉO GREMEN MIMARY DE OLIVEIRA

Aplicação do sequenciamento de nova geração no diagnóstico molecular de cardiomiopatia hipertrófica

Dissertação apresentada à Faculdade de Medicina da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Programa de Ciências Médicas

Área de concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo

Orientador: Dr. Alexandre da Costa Pereira

São Paulo

2015

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

©reprodução autorizada pelo autor

Oliveira, Théo Gremen Mimary de

Aplicação do sequenciamento de nova geração no diagnóstico molecular de cardiomiopatia hipertrófica / Théo Gremen Mimary de Oliveira. -- São Paulo, 2015.

Dissertação(mestrado)--Faculdade de Medicina da Universidade de São Paulo. Programa de Ciências Médicas. Área de concentração: Distúrbios Genéticos de Desenvolvimento e Metabolismo.

Orientador: Alexandre da Costa Pereira.

Descritores: 1.Cardiomiopatia hipertrófica 2.Genética médica 3.Mutação
4.Sequenciamento de nucleotídeos em larga escala 5.Biologia computacional
6.Técnicas de diagnóstico molecular

USP/FM/DBD-191/15

Dedicatória

Dedico este trabalho aos meus pais, por me darem a vida e me ensinarem tanto a vivê-la como a admirá-la. Não há nada no que eu sou, que eu não deva a vocês.

Agradecimentos

Ao meu orientador Alexandre Pereira, por colocar em minhas mãos a responsabilidade de iniciar esse projeto e acreditar na minha capacidade em desenvolvê-lo. Ao Miguel Mitne Neto e a Louise Teixeira, por todo auxílio que me foi prestado durante o dia-a-dia laboratorial e por todo o incentivo que me deram. A Julia Marsiglia, por me “arrastar” para o programa de hipertrófica e ter me dado todo suporte na elaboração do projeto inicial e durante todo o percurso até aqui. Vocês são os principais responsáveis por tudo o que eu aprendi nesses últimos três anos.

Ao Grupo Fleury, por toda infraestrutura cedida durante esse projeto e a equipe do setor de pesquisa e desenvolvimento. Vocês foram mais que fundamentais!

A Capes, ao FINEP e a Faculdade de Medicina da USP pelo financiamento e pela oportunidade de desenvolver esse projeto.

A todos os meus amigos do Laboratório de Genética e Cardiologia Molecular: Ju, Cinthia, Fanny, Gabi, Vaquero, Kaká, Diogo, Michelle Sabrina, pessoal do grupo do Paulo Caleb, meninas da Proteômica, meninas do HF, pessoal da Bioinfo, do Pós-doc, do Laboratório geral, da cultura de células, meninas da recepção e da secretaria. O cotidiano com vocês é algo que me estimula a levantar da cama todos os dias.

À toda minha família, simplesmente por vocês serem quem são e por me ensinarem a sempre buscar o conhecimento das coisas, esteja ele onde estiver. Minha alma de biólogo nada mais é do que responsabilidade de vocês! Por todas as vezes que eu ouvi da

minha avó Júlia que todo ser vivo, por menor que seja, tem sua função na natureza. Por todas as vezes em que minha tia Selma me contou (em plenas férias de verão) como Thomas Morgan usou aquelas mosquinhas da banana para estudar a hereditariedade, como Marie Curie descobriu o raio-X e como as conchas iam parar na beira da praia. Por todos os casos misteriosos da medicina que a Carol e meu pai discutiram em muitos almoços de domingo, sempre interrompendo a discussão para sanar minha faminta curiosidade. Por sempre me manterem cercado de animais, sejam eles peixes, répteis, aves ou mamíferos. E, com certeza, por muitos outros momentos e motivos que agora me escapam da memória, mas que fazem parte de mim e sempre farão.

Ao meu irmão Tales, pela cumplicidade musical de todos esses anos.

A Ariane, pelo companheirismo de sempre, pelo estímulo, pela compreensão e por tornar os meus dias mais leves. A Sandra, Dam e D. Amélia, pelo carinho.

Aos meus amigos da faculdade e a todos os membros da Nilson's Farm Corporation. Ter vocês como amigos é o maior presente que eu poderia ter recebido na vida.

*“O que eu sou, eu sou em par.
Não cheguei sozinho.”*

(Lenine – Castanho)

Sumário

Lista de Siglas

Lista de Tabelas

Lista de Figuras

Resumo

Abstract

1.	INTRODUÇÃO	1
1.1	Cardiomiopatia hipertrófica.....	1
1.1.1	Aspectos clínicos e genéticos.....	1
1.1.2	Importância do diagnóstico molecular	4
1.2	Sequenciamento de nova geração.....	6
1.2.1	Surgimento e aplicações nas doenças mendelianas	6
1.2.2	Plataforma Ion Torrent PGM™	10
2.	OBJETIVOS.....	14
2.1	Objetivo geral	14
2.2	Objetivos específicos.....	14
3.	MATERIAL E MÉTODOS	15
3.1	Extração de DNA	15
3.2	Desenho do Painel Genético	15
3.3	Composição da casuística	16
3.4	Desenho experimental	17
3.5	Processo de enriquecimento	18
3.6	Preparo de <i>template</i> e sequenciamento.....	19
3.7	Análises de Bioinformática	20
3.8	Processo de validação e avaliação analítica	22
3.9	Análise de ganho diagnóstico.....	23
4.	RESULTADOS	26
4.1	Desempenho Geral do Sequenciamento	26
4.2	Validação analítica	27
4.2.1	Cobertura de Regiões Alvo.....	27
4.2.2	Sensibilidade e Especificidade.....	29
4.2.3	VPP, TFD e Reprodutibilidade	32
4.2.4	Investigação de variantes não encontradas e discrepantes.....	36
4.3	Ganho diagnóstico	39
5.	DISCUSSÃO	46
5.1	Desempenho analítico do <i>pipeline</i>	46
5.2	Ganho diagnóstico	51
6.	CONCLUSÕES	56
7.	ANEXOS	57
8.	REFERÊNCIAS BIBLIOGRÁFICAS	58

Lista de Siglas

BAM	<i>Binary Alignment Map</i>
CADD	<i>Combined Annotation Dependent Depletion</i>
CD	Cardiomiopatia dilatada
CH	Cardiomiopatia hipertrófica
DAVD	Displasia Arritmogênica de Ventrículo Direito
dbSNP	<i>Single Nucleotide Polymorphisms database</i>
EVS	<i>Exome Variant Server</i>
ExAC	<i>Exome Aggregation Consortium</i>
FAV	Frequência do alelo variante
FP	Falso positivo
HGMD	<i>Human Genome Mutation Database</i>
MAF	<i>Minor Allele Frequency</i>
NC	Não Concordante
SNG	Sequenciamento de Nova Geração
TFD	Taxa de falsas descobertas
VCF	<i>Variant Call Format</i>
VE	Ventrículo esquerdo
VP	Verdadeiro positivo
VPP	Valor preditivo positivo

Lista de Tabelas

Tabela 1: Genes e suas respectivas frequências de associação com CH de acordo com o banco de dados Online Mendelian Inheritance in Man (OMIM)... 2	2
Tabela 2: Corridas realizadas e a respetiva distribuição das amostras ao longo destas. Amostras negativas incluídas nas corridas de validação também tiveram seus resultados finais utilizados nas análises de ganho diagnóstico. . 18	18
Tabela 3: Desempenho do sequenciamento na plataforma Ion Torrent PGM™ 26	26
Tabela 4: Porcentagem de cobertura do painel analisada em 7 pontos diferentes de cobertura. 28	28
Tabela 5: Valores de sensibilidade e especificidade da chamada de variantes nos genes MYH7, MYBPC3 e TNNT2, com as amostras da casuística de CH previamente processadas no Painel CH-Sanger..... 29	29
Tabela 6: Mutações patogênicas presentes nas amostras da casuística de CH e suas respectivas confirmações. 30	30
Tabela 7: Resultados da validação analítica inter-ensaio realizada com a amostra NA12878. Abreviações: Não concordante (NC); Valor Preditivo Positivo (VPP); Taxa de Falsas Descobertas (TFD); Reprodutibilidade (Rep) e Intervalo de Confiança (IC)..... 34	34
Tabela 8: Resultados da validação analítica intra-ensaio realizada com a amostra NA12878. Abreviações: Não concordante (NC); Valor Preditivo Positivo (VPP); Taxa de Falsas Descobertas (TFD); Reprodutibilidade (Rep) e Intervalo de Confiança (IC)..... 35	35
Tabela 9: Variantes perdidas no processo de validação com as amostras da casuística de CH. 36	36
Tabela 10: Análise das variantes não encontradas e discrepantes no experimento inter-ensaio com a amostra NA12878. 37	37
Tabela 11: Análise das variantes não encontradas e discrepantes no experimento intra-ensaio com a amostra NA12878. 38	38

Tabela 12: Alterações encontradas nas amostras reclassificadas como positivas. Abreviações: Cardiomiopatia hipertrófica (CH); Amiloidose (Amil.); Síndrome de Noonan (Noonan); PROVEAN (Pr); SIFT (S); PolyPhen-2 (P2). 40

Tabela 13: Alterações encontradas nas amostras reclassificadas como inconclusivas. Abreviações: Cardiomiopatia dilatada (CD); Cardiomiopatia hipertrófica (CH); Displasia arritmogênica de ventrículo esquerdo (DAVD); Morte súbita infantil (MSI); PROVEAN (Pr); SIFT (S); PolyPhen-2 (P2). 41

Lista de Figuras

Figura 1: Queda no custo de sequenciamento de DNA nos últimos 13 anos (em dólares/genoma humano).	7
Figura 2: Etapas envolvidas no processo de sequenciamento na plataforma Ion Torrent PGM™	11
Figura 3: Esquema ilustrando os processos de preparo, sequenciamento e análises de bioinformática aos quais as amostras eram submetidas. Os quadrados com borda serrilhada indicam etapas envolvidas no processo de validação do <i>pipeline</i>	20
Figura 4: Algoritmo de análise para alterações encontradas nos resultados de SNG.	25
Figura 5: Distribuição dos valores de cobertura média das 1754 regiões alvo. A barra preta indica as regiões com cobertura média abaixo de 10x.	27
Figura 6: Distribuição de cobertura das amostras envolvidas no processo de validação. Os grupos com cores intercaladas indicam quais amostras foram processadas dentro da mesma corrida.	28
Figura 7: Análise de reprodutibilidade entre as replicatas da amostra NA12878, nas perspectivas inter-ensaio (Painel A) e intra-ensaio (Painel B). Ao todo 134 variantes constavam na tabela benchmark para serem confirmadas.....	33
Figura 8: Proporção de reclassificação das 72 amostras negativas sequenciadas.	39
Figura 9: Frequência de aparecimento de alterações nos 19 genes com mutação detectada nas amostras positivas.	42
Figura 10: Frequência de aparecimento de alterações nos 16 genes com mutação detectada nas amostras inconclusivas.	45

Resumo

Oliveira, TGM. *Aplicação do sequenciamento de nova geração no diagnóstico molecular de cardiomiopatia hipertrófica* [dissertação]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2015

Introdução: A cardiomiopatia hipertrófica (CH) é uma doença cardíaca estrutural primária, caracterizada por hipertrofia do ventrículo esquerdo, sem dilatação, geralmente assimétrica e predominantemente septal. Na população geral a prevalência estimada da CH é de 0,2% (1:500), correspondendo a 0,5% de todas as cardiopatias. Atualmente estão descritas mais de 1400 mutações associadas à CH em 20 genes relacionados com os miofilamentos do sarcômero, o disco-Z e o transporte de cálcio, sendo que os três mais associados são os genes *MYH7*, *MYBPC3* e *TNNI2*, responsáveis por 50% dos casos com diagnóstico molecular positivo no Brasil. Dessa forma, o advento de novas tecnologias de sequenciamento de DNA de alta performance promete revolucionar o diagnóstico molecular, tornando mais rápida e barata a identificação de alterações genéticas, impactando positivamente na custo-efetividade do manejo diagnóstico e terapêutico de pacientes e famílias com o diagnóstico de CH. **Materiais e Métodos:** Noventa e uma amostras de uma casuística de pacientes não relacionados, portadores de CH com diagnóstico molecular prévio para os 3 genes mais associados (19 positivas e 72 negativas) foram utilizadas juntamente com uma amostra referência do HapMap (NA12878) na validação de um *pipeline* proposto para a identificação de alterações genéticas em um painel com 74 genes associados à cardiomiopatias hereditárias, utilizando a plataforma Ion Torrent PGM. A etapa de chamada de variantes foi testada em dois limiares diferentes de cobertura de sequenciamento (30x e 10x) e três limiares de frequência de alelo variante (35%, 25% e 20%). A amostra NA12878 foi utilizada na aferição de valores de reprodutibilidade intra e inter-ensaio. As amostras da casuística de CH com diagnóstico molecular prévio negativo foram utilizadas na análise de ganho diagnóstico. Eram consideradas alterações potencialmente patogênicas aquelas que apresentassem associação prévia com CH ou classificação deletéria em dois de três algoritmos de predição de impacto funcional (PROVEAN, SIFT, PolyPhen2) e $MAF < 0,01$, se disponível. **Resultados:** A plataforma de sequenciamento utilizada apresentou desempenho aceitável, gerando em média $165,9 \pm 13,1$ Mb, com um valor médio de $146,9 \pm 11,54$ Mb acima de $PhredQ \geq 20$, por amostra. O valor médio de cobertura de sequenciamento por amostra foi de $250 \pm 23,94x$, com 95,2% das regiões alvo cobertas pelo menos 10x. A sensibilidade máxima observada para SNVs foi de 96,7% enquanto que para InDels foi 28,5%. Os valores de reprodutibilidade inter e intra-ensaio de 89,5% e 87,3%, respectivamente. Das 72 amostras negativas, 35 puderam ser reclassificadas como positivas, sendo que os dois genes com mais ocorrências de alterações genéticas foram *FLNC* e *TRIM63*, ambos já relacionados com CH. Vinte e duas amostras foram reclassificadas como inconclusivas e 15 permaneceram negativas. O ganho diagnóstico foi de 21,5%. **Conclusões:** A plataforma Ion Torrent PGM apresenta potencial no sequenciamento de genes relacionados à cardiomiopatias hereditárias e o *pipeline* validado mostrou valores analíticos praticáveis em uma rotina diagnóstica. A utilização do painel genético ampliado se mostrou viável na detecção de alterações genéticas, propiciando uma boa margem de ganho diagnóstico em comparação com o sequenciamento apenas dos três genes mais associados à CH.

Descritores: Cardiomiopatia hipertrófica; genética médica; mutação; sequenciamento de nucleotídeos em larga escala; biologia computacional; técnicas de diagnóstico molecular.

Abstract

Oliveira, TGM. *Application of next-generation sequencing in the molecular diagnostics of hypertrophic cardiomyopathy* [dissertation]. "São Paulo: Faculdade de Medicina, Universidade de São Paulo"; 2015

Introduction: Hypertrophic cardiomyopathy (HCM) is a primary cardiac disease, mainly characterized by unexplained left ventricle hypertrophy, in the absence of dilatation, usually asymmetric and predominantly septal. The estimated prevalence is 1:500 individuals in the general population, corresponding for 0.5% of all cardiac diseases. Up to now, more than 1400 mutations are associated with HCM in 20 genes related with sarcomeric myofibrils, Z-disc and calcium homeostasis, wherein the 3 most associated genes are *MYH7*, *MYBPC3* and *TNNT2*, accounting for 50% of cases with positive molecular diagnostics in Brazil. Thus, the advent of new high throughput DNA sequencing technologies promise to revolutionize the use of molecular diagnostics, turning the identification of genetic mutations in a fast and cheap practice, increasing the cost-effectiveness of diagnostic and treatment of patients and families with HCM.

Materials and Methods: Ninety one samples from an HCM casuistic of unrelated individuals with previous molecular diagnostics for the three most HCM-associated genes (19 positives and 72 negatives) were processed along with a reference HapMap sample (NA12878) in the validation process of a pipeline proposed for the detection of genetic alterations in a genetic panel composed of 74 genes associated with inherited cardiomyopathies, using Ion Torrent PGM platform. The variant call step was tested for two cutoffs of sequencing coverage (30x and 10x) and three cutoffs of variant allele frequency (35%, 25% and 20%). The sample NA12878 was used in the assessment of intra and inter-assay reproducibility. Negative samples from the HCM casuistic were used in the assessment of diagnostic yield. Variants were considered potentially pathogenic if previously described as associated with HCM or if presenting a deleterious score in at least two of three impact prediction algorithms tested (PROVEAN, SIFT and PolyPhen-2) and $MAF < 0.01$, if available. **Results:** The chosen next-generation sequencing platform presented an acceptable performance, with a mean throughput of $165,9 \pm 13,1$ Mb, with a mean value of $146,9 \pm 11,54$ Mb above $PhredQ \geq 20$. Mean sequencing coverage was $250 \pm 23,94x$, wherein 95.2% of target bases were covered at least 10x. Maximum achieved sensitivity for SNVs was 96.7% while for InDels was 28.5%. Both values of inter and intra-assay reproducibility were 89.5% and 87.3%, respectively. Of all 72 negative samples, 35 were reclassified as positive with the two most frequently mutated genes being *FLNC* and *TRIM63*, both already associated with HCM. Twenty two samples were reclassified as inconclusive and 15 remained negatives. Diagnostic yield was 21.5%. **Conclusions:** Ion Torrent PGM platform presented a feasible potential for the sequencing of inherited cardiomyopathies-associated genes and the designed pipeline presented reliable analytical values for diagnostic use. The expanded panel proved to be a good strategy for the detection of genetic alteration providing a good value of diagnostic yield in comparison with the sequencing of the three most HCM-associated genes alone.

Descriptors: cardiomyopathy, hypertrophic; genetics, medical; mutation; High-throughput nucleotide sequencing; Computational biology; Molecular diagnostics techniques.

1. INTRODUÇÃO

1.1 Cardiomiopatia hipertrófica

1.1.1 Aspectos clínicos e genéticos

A cardiomiopatia hipertrófica (CH) é uma cardiopatia estrutural primária, cuja principal característica é a hipertrofia inexplicada do ventrículo esquerdo (VE), de predominância geralmente septal, na ausência de dilatação ventricular e qualquer outra condição cardíaca ou sistêmica que possa causar hipertrofia miocárdica secundária como hipertensão, estenose aórtica ou amiloidose. De todas as cardiomiopatias hereditárias é a mais frequente, apresentando uma prevalência de 1:500 (0,2%) na população geral. É uma doença de espectro global, tendo sido relatada em mais de 50 países, em todos os continentes, afetando igualmente indivíduos de diferentes sexos e etnias¹.

Alguns dos sintomas mais observados são dispneia, angina, palpitações e síncope, porém a manifestação clínica da CH é altamente heterogênea, com a apresentação e a progressão dos sintomas variando consideravelmente entre indivíduos. Enquanto alguns apresentam já de início sinais de hipertrofia acentuada, insuficiência cardíaca e arritmias, outros podem permanecer assintomáticos durante toda a vida, o que confere à CH uma expressividade variável².

A ocorrência é geralmente familiar apresentando, na maioria dos casos, padrão de herança mendeliano autossômico dominante, porém há descrição de casos esporádicos, bem como de casos autossômicos recessivos^{3,4}. Em 1989 foi descrita a primeira associação de um marcador cromossômico com a transmissão familiar, indicando a banda q1 do cromossomo 14 como o *locus* do

possível gene mutado, o qual posteriormente seria conhecido como a localização do gene codificante da cadeia pesada da β -miosina (*MYH7*)⁵. Desde então, 20 genes já foram associados⁶ (Tabela 1), sendo que os três genes mais implicados são os genes *MYH7*, *MYBPC3* (Proteína C de ligação da miosina) e *TNNT2* (Troponina T), todos codificantes de miofibrilas musculares, o que levou a denominação inicial de “doença do sarcômero” à CH.

Tabela 1: Genes e suas respectivas frequências de associação com CH de acordo com o banco de dados Online Mendelian Inheritance in Man (OMIM).

Gene	Proteína	Locus Cromossômico	Frequência
Sarcômero			
<i>TNNT2</i>	Troponina T	1q32.1	3-5%
<i>TTN</i>	Titina	2q31.2	<1%
<i>MYL3</i>	Cadeia Leve da Miosina Essencial	3p21.31	<1%
<i>TNNC1</i>	Troponina C	3p21.1	<1%
<i>MYBPC3</i>	Proteína C de Ligação da Miosina	11p11.2	25-35%
<i>MYL2</i>	Cadeia Leve da Miosina Regulatória	12q24.11	<2%
<i>MYH7</i>	Cadeia pesada da β -miosina	14q11.2	25-35%
<i>MYH6</i>	Cadeia pesada da α -miosina	14q11.2	<1%
<i>TPM1</i>	α -Tropomiosina	15q22.2	<5%
<i>ACTC1</i>	α -Actina cardíaca	15q14	<1%
<i>TNNI3</i>	Troponina I	19q13.42	<5%
Disco-Z			
<i>NEXN</i>	Nexilina	1p31.1	Rara
<i>MYOZ2</i>	Miozenina	4q26	<1%
<i>VCL</i>	Vinculina	10q22.2	<1%
<i>CSRP3</i>	Proteína muscular LIM	11p15.1	<1%
Controle de Cálcio			
<i>PLN</i>	Fosfolambam	6q22.31	<1%
<i>CALR3</i>	Calreticulina	9p13.11	<1%
<i>JPH2</i>	Junctofilina	20q13.12	<1%
Outros			
<i>CAV3</i>	Caveolina	3p25.3	Rara
<i>MYLK2</i>	Quinase da cadeia leve da miosina	20q11.21	Rara

Dados internacionais indicam que alterações nesses genes podem chegar a 75% dos casos com diagnóstico molecular positivo^{1,7}. No Brasil, especificamente, tal envolvimento representa aproximadamente 50% dos casos⁸. Além de genes sarcoméricos, estão envolvidos também genes relacionados à estrutura do disco-Z e ao controle de cálcio no músculo cardíaco e, até o momento, mais de 1400 mutações genéticas já foram descritas como causais ou associadas à CH, indicando tanto heterogeneidade gênica como alélica para a doença^{1,9}.

Mesmo ocorrendo em núcleos familiares a CH apresenta penetrância altamente variável, dado que indivíduos de uma mesma família portadores de um mesmo alelo alterado podem apresentar diferentes idades de diagnóstico, apresentação e evolução de quadro clínico. O aparecimento de hipertrofia acontece geralmente na adolescência ou no início da idade adulta, mas pode estar presente ao nascimento ou aparecer mais tardiamente em indivíduos com mais de 50 anos, porém a penetrância raramente atinge 100% e, em uma parcela dos indivíduos portadores de mutação, o fenótipo pode permanecer ausente (denominados “genótipo positivo – fenótipo negativo”)¹⁰. Frente a essa característica alguns autores sugerem que apenas as alterações causais não são o único determinante na variabilidade da penetrância e que alterações concomitantes em outros genes podem agir como moduladores fenotípicos^{1,11}.

A hipertrofia e o desarranjo das fibras sarcoméricas são características histológicas marcantes da doença. O desarranjo de miofibrilas parece estar intimamente associado com o aparecimento de fibrose, a qual, por sua vez, ocorre por conta da expansão da matriz intersticial decorrente da morte

precoce de cardiomiócitos, seja por apoptose prematura ou por isquemia miocárdica⁷. Tal cenário contribui principalmente para o desenvolvimento de insuficiência cardíaca e no surgimento de arritmias ventriculares, sendo a última um conhecido fator de predisposição à morte súbita a qual, apesar de não muito frequente, é um preocupante desfecho da CH, uma vez que ocorre com mais frequência em indivíduos jovens e é apontada como frequente causa de morte em atletas¹.

1.1.2 Importância do diagnóstico molecular

A utilização de testes genéticos em casos de CH tem aumentado nos últimos anos em decorrência de benefícios observados tanto no contexto individual como no familiar. A princípio, nas situações em que o diagnóstico clínico é uma certeza, o estabelecimento do defeito molecular reserva-se apenas a confirmação diagnóstica. Ainda assim, o diagnóstico molecular pode contribuir como uma evidência em casos em que a confirmação clínica é incerta como, por exemplo, na existência de hipertrofia limítrofe do VE, em casos de hipertrofia identificada em atletas (também denominado “coração de atleta”) e ainda em casos de suspeita de hipertrofia por doenças de depósito.

Porém, uma das principais aplicações da análise genética é o diagnóstico pré-clínico em indivíduos membros de famílias com histórico da doença. Uma vez identificada presença de mutação em heterozigose, cada membro apresenta 50% de chances de herdá-la, o que faz do diagnóstico molecular uma ferramenta importante na identificação de indivíduos em risco, principalmente na existência de histórico de morte súbita. Em alguns casos, esse pode ser a primeira manifestação da doença, sendo que ocorre em 1 a

2% das crianças e adolescentes e em 0,5 a 1% dos jovens adultos com mutações causais identificadas¹².

Contudo, deve-se entender que o diagnóstico molecular positivo, especialmente em indivíduos assintomáticos, não significa confirmação clínica da doença e sim um risco aumentado ao desenvolvimento da mesma. Quando se tem uma mutação identificada, o rastreamento de familiares se torna uma estratégia muito importante, pois pode incentivar um acompanhamento adequado dos indivíduos que possuem a mutação além de assegurar aos familiares que não a possuem, a ausência de risco de desenvolvimento da doença. Sabe-se inclusive que mutações em genes sarcoméricos estão mais associados com idade precoce de diagnóstico, hipertrofia acentuada, histórico de CH e morte súbita na família^{13,14}. Indivíduos portadores de alterações nesses genes, principalmente quando no contexto de uma família com CH, seriam candidatos a um controle mais rígido de fatores de risco de desenvolvimento e progressão da doença, assim como de uma monitorização médica mais rigorosa. Ainda, pacientes portadores de mais de uma alteração genética, seja no mesmo gene (heterozigose composta) ou em genes diferentes (duplo heterozigoto) também estão mais propensos a apresentar um fenótipo clínico mais grave¹⁵.

A identificação de alterações genéticas patogênicas é um fator que impacta também na custo-efetividade do processo de rastreamento familiar. O rastreamento genético de pacientes com CH e seus familiares é a estratégia mais custo efetiva quando comparada ao rastreamento clínico isolado, uma vez que passam a ser seguidos de forma mais próxima apenas os indivíduos com

mutação confirmada¹⁶, diminuindo gastos com consultas médicas e exames laboratoriais por parte dos não portadores. Nesse contexto, o Sequenciamento de Nova Geração (SNG) se apresenta como uma alternativa promissora, pois permite a análise de conjuntos de genes de forma rápida, objetiva e mais barata do que o método clássico de sequenciamento, tornando o rastreamento molecular uma abordagem exequível no cuidado familiar.

1.2 Sequenciamento de nova geração

1.2.1 Surgimento e aplicações nas doenças mendelianas

Em 1977, o inglês Frederick Sanger et al.¹⁷ publicaram o método de sequenciamento por terminação de cadeia (popularmente conhecido na comunidade científica como “sequenciamento de Sanger”), o qual rapidamente se difundiu e se tornou o método de sequenciamento mais utilizado e aprimorado nas três décadas seguintes. O Projeto Genoma Humano (PGH), primeira iniciativa internacional de sequenciamento em larga escala finalizado em 2003, utilizou o método de Sanger como técnica principal, o que contribuiu para o projeto ter duração total de 13 anos ao custo de US\$2,7 bilhões¹⁸. Dessa forma, após a nítida percepção das limitações dessa técnica em projetos de alta demanda, novos esforços surgiram para criação de técnicas mais rápidas e baratas para o sequenciamento de genomas.

A partir de 2005, diversas plataformas de SNG surgiram no mercado, rapidamente tornando o sequenciamento de larga escala mais rápido, barato e acessível¹⁹. Nos últimos 12 anos o custo do sequenciamento genético apresentou queda exponencial (Figura 1), tornando o SNG uma prática cotidiana em diversos laboratórios de pesquisa e diagnóstico no mundo todo,

expandindo o escopo de descobertas tanto no nível individual como populacional. Ao mesmo tempo em que a geração de dados de SNG cresceu, uma série de ferramentas computacionais foram desenvolvidas com intenção de acelerar a capacidade de análise desses dados, aumentando o campo de intersecção entre a biologia molecular e informática, assim consolidando a área de atuação conhecida hoje como bioinformática. Tais processos computacionais levaram ao surgimento de programas de encadeamento de dados (*pipelines*) específicos para SNG, sendo peças fundamentais no tipo e na qualidade da análise desejada²⁰.

Figura 1: Queda no custo de sequenciamento de DNA no período de 2002 a 2013 (em dólares/genoma humano)²¹.

A alta capacidade de geração de dados, somada ao constante desenvolvimento de *pipelines* levou o SNG a apresentar um largo espectro de aplicações na pesquisa científica, sendo utilizado não apenas no estudo de genomas completos, como também em conjuntos específicos de genes,

regiões regulatórias do genoma, sequenciamento de RNA, entre outras técnicas passíveis de aplicação tanto para seres procariotos como eucariotos²². No que diz respeito aos cuidados com a saúde humana, e mais especificamente às doenças de herança mendeliana, o SNG tem como principais aplicações o sequenciamento de exoma completo²³, referente à todas as regiões codificantes de proteínas do genoma humano (aproximadamente 2%) e o sequenciamento alvo²⁴, o qual limita-se apenas a alguns grupos de genes (também chamados de “painéis genéticos”) já envolvidos com a doença de interesse.

Ambas as estratégias apresentam a mesma finalidade: encontrar alterações em regiões codificantes para as quais se possa atribuir causalidade direta a um fenótipo observado. O uso do sequenciamento de exoma completo é uma prática que tem sido amplamente adotada, mostrando resultados efetivos em relação a descoberta de novos genes delineando doenças mendelianas^{25,26}. Porém, quando se trata de aplicações clínicas mais diretas, algumas opiniões ainda parecem favorecer o uso de painéis genéticos principalmente por duas razões: maior objetividade e rapidez na interpretação e no retorno de resultados e melhor desempenho em relação a cobertura de sequenciamento dos genes de interesse^{24,27}. Em relação às cardiomiopatias hereditárias, tal preferência por painéis genéticos parece se confirmar na literatura, mostrando que a aplicação dessa abordagem é viável num contexto de diagnóstico, sendo capaz de cobrir os genes mais importantes e ainda propiciar novas descobertas²⁸⁻³¹.

Um dos principais desafios em relação à descoberta de novas alterações genéticas é a interpretação do impacto dessas na funcionalidade de um gene e, conseqüentemente, a atribuição de causalidade. Atualmente, duas estratégias muito adotadas no processo de interpretação de variantes são: uso de algoritmos computacionais de predição de impacto funcional (também conhecida como abordagem *in silico*) e análise de frequência populacional da alteração. A primeira se baseia no fato de que regiões do genoma mais conservadas num espectro filogenético seriam menos tolerantes a alterações genéticas, pois ao longo da cadeia evolutiva foram submetidas a uma rígida seleção negativa, e uma mutação nessas regiões teria um maior impacto do que em regiões genômicas variáveis. Assim, algoritmos analisam o perfil filogenético da região alterada e tentam predizer qual o peso de uma eventual alteração genética observada naquela região, em função do seu grau de conservação evolutiva³².

Já a segunda estratégia consiste na busca pela alteração de interesse em bancos de dados populacionais, com o intuito de se inferir a frequência da alteração encontrada na população geral. Com a disponibilização dos resultados de sequenciamento de genoma e exoma completo de diversos indivíduos, como ocorre nos projetos “1000 Genomas”³³ e “*Exome Sequencing Project* (ESP)”³⁴, torna-se possível calcular a frequência populacional de determinadas alterações. Dado que doenças mendelianas acontecem geralmente em uma prevalência baixa na população (abaixo de 1%), uma possível alteração causal tende a apresentar uma frequência igual ou menor.

Apesar dessas estratégias já serem de uso frequente por muitos laboratórios, a interpretação do impacto de variantes ainda não é um processo linear e muitos pontos ainda permanecem sujeitos a subjetividade dos pesquisadores. Na tentativa de se diminuir vieses, algumas instituições divulgam permanentemente diretrizes de interpretação, como é o caso do Colégio Americano de Genética Médica (*American College of Medical Genetics* ou ACMG)³⁵ nas quais são definidas coordenadas para coleta de evidências de malignidade ou benignidade de uma alteração. Da mesma forma, sugere-se que *pipelines* utilizados em aplicações clínicas de SNG passem por uma extensa validação antes de seu uso rotineiro, assim permitindo a aferição de parâmetros analíticos (sensibilidade, especificidade, reprodutibilidade, entre outros) e possíveis pontos a serem otimizados^{36,37}.

1.2.2 Plataforma Ion Torrent PGM™

Lançada em 2011, a plataforma Ion Torrent PGM™ (Life Technologies, Thermo Scientific) foi apresentada como um sequenciador de bancada (*benchtop*), direcionado para sequenciamento de pequenos genomas e/ou grupos de genes. Foi a primeira a apresentar uma química de sequenciamento totalmente independente de moléculas modificadas, sem uso de nucleotídeos fluorescentes, o que dispensa o uso de aparatos ópticos, tornando tanto o equipamento como seus reagentes mais baratos.

O fluxo de bancada completo inclui três etapas: 1) Construção da biblioteca, 2) Preparo de *template* e 3) Sequenciamento em chip de semicondutor (Figura 2). A construção de biblioteca inicia-se com a fragmentação do DNA de interesse, seja por via enzimática ou mecânica. Em

seguida, os fragmentos recebem em suas extremidades dois adaptadores de DNA com sequência conhecida (denominados A e P1), os quais são essenciais para garantir o seu sequenciamento (Figura 2, Painel 1). Caso haja intenção de se processar mais de uma amostra por corrida de sequenciamento, são utilizados juntos a esses adaptadores sequências denominadas *barcodes*, as quais são específicas para cada amostra, permitindo que no momento da análise cada amostra seja devidamente separada. Geralmente nesse passo, também se realiza uma seleção de tamanho de fragmentos para que a maioria dos fragmentos a serem sequenciados estejam entre 200 e 400pb.

Figura 2: Etapas envolvidas no processo de sequenciamento na plataforma Ion Torrent PGM™.

Preparados os fragmentos, estes são encaminhados para o preparo de *template*, cuja principal etapa é a PCR em emulsão (emPCR), a qual consiste na amplificação clonal dos fragmentos dentro de micelas geradas em uma solução emulsionada. Dentro de cada microrreator há um grânulo magnético (*bead*) com sequências de DNA na superfície, no qual um fragmento se acoplará através de seu adaptador P1 (Figura 2, Painel 2). A amplificação clonal dos fragmentos ocorre através do pareamento dos sítios de ligação dos *primers* presentes nos adaptadores e, conseqüentemente, com a extensão de uma nova fita de DNA. Terminado o tempo da emPCR, uma fase de enriquecimento separará os *beads* acoplados à fragmentos dos não acoplados, e aqueles selecionados serão inseridos no chip semicondutor e encaminhados para o equipamento, onde acontecerá a reação de sequenciamento.

Na última etapa, a amostra já devidamente preparada é inserida em um chip semicondutor, o qual é constituído por poços de diâmetro micrométrico nos quais as *beads* com DNA serão depositados para sequenciamento. Cada poço recebe uma *bead* e embaixo de cada um há um aparato semicondutor capaz de detectar variações elétricas. Assim, o equipamento passa a fornecer às *beads* os quatro nucleotídeos (A,C, T ou G) soltos, porém sempre um de cada vez, nunca misturados. Caso haja a incorporação de algum nucleotídeo em algum fragmento de DNA, a ligação fosfodiéster gerada liberará íons de hidrogênio no meio, alterando o pH e fazendo com que essa variação seja detectada pelos semicondutores. Assim, sempre que houver o fornecimento de um nucleotídeo seguido de variação de pH, significa que aquela base foi incorporada, permitindo a geração de sequências ao final do processo .

O tempo total de sequenciamento pode chegar a 5 horas, porém ao final do processo podem ser lidas mais de um bilhão de pares de bases (>1Gb), fazendo dessa plataforma uma técnica extremamente eficaz no que diz respeito ao consumo de tempo na geração de dados brutos. Assim, em virtude da condição genética apresentada acima, com ao menos 20 genes passíveis de investigação para CH, somado ao fator eficiência do SNG, este trabalho tem por finalidade a aplicação do sequenciamento massivo de DNA em pacientes portadores da doença, no intuito de se demonstrar a eficiência dessa técnica no diagnóstico molecular, bem como determinar o ganho diagnóstico obtido com o uso de painéis genéticos ampliados.

2. OBJETIVOS

2.1 Objetivo geral

O objetivo principal deste trabalho foi a aplicação do sequenciamento de nova geração no diagnóstico molecular de cardiomiopatia hipertrófica, com a utilização do método de captura em solução.

2.2 Objetivos específicos

- Padronizar e otimizar o sequenciamento de próxima geração da plataforma Ion Torrent PGM[™];
- Testar a aplicabilidade do método de captura em solução no diagnóstico de CH, através do sequenciamento de nova geração;
- Validar a análise de 74 genes relacionados à CH e a fenocópias em uma população já com diagnóstico molecular prévio realizado pelo método de sequenciamento tradicional (Sanger), e uma amostra referência do projeto HapMap (NA12878);
- Montagem e validação de um *pipeline* computacional para análise de resultados de sequenciamento de nova geração;
- Análise do ganho diagnóstico em um painel ampliado de genes relacionados com cardiomiopatias hereditárias, em pacientes com resultado de diagnóstico molecular prévio negativo.

3. MATERIAL E MÉTODOS

3.1 Extração de DNA

Foram coletados 4mL de sangue periférico em tubo de EDTA, de cada indivíduo. A extração de DNA foi realizada como descrito por Miller et.al³⁸. O sangue foi hemolisado com tampão contendo NH₄Cl 0,144M e NH₄CO₃ 0,001M. Em seguida os leucócitos foram lisados (soluções Tris 0,01M, NaCl 0,4M, EDTA 0,002M em pH 8,0 e EDTA 0,5M, SDS 10% em pH 8,0), o DNA foi precipitado (solução NaCl 6M) e ressuspenso em TE (Tris-HCL 10mM, EDTA 1mM em pH 8,0). Uma primeira aferição da concentração da solução de DNA obtida foi determinada com leitura em espectrofotômetro a 260nm. A solução de DNA era diluída em água (40ng/mL) para uso e armazenada a -20°C.

3.2 Desenho do Painel Genético

A tecnologia utilizada para o enriquecimento dos 74 genes alvo (Anexo A) foi o kit *Haloplex Target Enrichment System* (Agilent Technologies). O desenho inicial das sondas foi feito a partir da plataforma virtual SureDesign, na qual foram indicados os genes de interesse (através das siglas do banco RefSeq) e a versão do genoma humano referência (hg19). Apenas regiões codificantes foram indicadas como alvo, mantendo 10pb de regiões intrônicas nas extremidades 3' e 5' de cada exon. Foram inseridos genes relacionados principalmente ao sarcômero, disco-z, citoesqueleto, desmossomo e controle de cálcio. Foram incluídos também genes relacionados com as doenças do espectro Noonan (*PTPN11*, *KRAS*, *HRAS*, *SOS1*, *RAF1* e *SPRED1*), bem como genes relacionados com doenças de depósito, tais como amiloidose cardíaca e doenças de Pompe e Fabry (*TTR*, *GAA* e *GLA*, respectivamente).

Ao final do processo, foram fornecidos pela plataforma virtual arquivos de conferência contendo as regiões gênicas encobertas pelo desenho e as regiões não cobertas (não passíveis de sequenciamento). O desenho final consistia de 1754 regiões alvo, abrangendo 376553pb. Apenas 1,08% de todas as bases alvo foram indicadas como “perdidas” e a cobertura total das regiões alvo foi de 98,92%. O valor médio de conteúdo GC das regiões alvo foi de 47,7%, com valores mínimos e máximos de 20,7% e 74,8%, respectivamente. O montante a ser sequenciado por amostra, recomendado pelo fabricante, foi de 181,36Mb para se atingir uma cobertura esperada de 200x.

3.3 Composição da casuística

Amostras de DNA foram selecionadas a partir de uma casuística previamente estudada⁸ de indivíduos não relacionados com diagnóstico clínico positivo para CH provindos do Instituto do Coração (InCor/HC-FMUSP). O diagnóstico clínico foi realizado por cardiologistas do ambulatório de cardiomiopatias do InCor e o diagnóstico molecular foi realizado através da técnica de sequenciamento *Sanger* para os 3 genes mais associados com CH: *MYH7*, *MYBPC3* e *TNNT2* (aqui denominado como “Painel CH-Sanger”). Um conjunto experimental foi composto com 91 amostras das quais 19 tinham diagnóstico molecular prévio positivo (aqui referidas como P1 a P19), com pelo menos uma mutação potencialmente patogênica em um dos três genes analisados, e 72 tinham diagnóstico molecular negativo (N1 a N72).

Uma amostra referência (NA12878) do projeto HapMap (*International HapMap Project*) foi selecionada para o processo de validação do *pipeline*. Essa amostra apresenta uma lista de variantes de alta resolução (*benchmark*),

originária da junção de resultados de análises dessa amostra em diversas plataformas de SNG e processos de bioinformática, sendo amplamente indicada para avaliações de performance analítica em processos de validação³⁹.

3.4 Desenho experimental

O desenho experimental contou com 20 corridas de sequenciamento (corridas 1 a 20), cada uma contendo 5 amostras, com exceção da corrida 15 que foi processada com 4 amostras (Tabela 2). As corridas 1 a 7, 11, 16, 17, 18 e 20 foram organizadas para conter amostras utilizadas no processo de validação do *pipeline*. Ao todo nessas corridas, foram sequenciadas 52 amostras (19 positivas e 33 negativas), as quais juntas somavam 285 alterações (3 InDels e 282 SNVs) a serem confirmadas pela técnica proposta. A amostra NA12878 foi submetida ao processo de enriquecimento em triplicata e utilizada para as avaliação de reprodutibilidade inter-ensaio nas corridas 16, 17 e 18 (NA12878_1, _2 e _3) e intra-ensaio na corrida 20 (NA12878_1.1, _2.1 e _3.1) sendo que dentro das regiões de interesse se encontravam 134 alterações (132 SNVs, uma inserção e uma deleção) a serem confirmadas. As amostras P11 e P13 foram resequenciadas para o processo de análise de reprodutibilidade inter-ensaio sendo sequenciadas na corrida 20 (P11.1 e P13.1), juntamente com as 3 triplicatas da amostra NA12878. O restante das corridas contou com 39 amostras negativas, as quais foram utilizadas para as análises de ganho diagnóstico, juntamente com as amostras negativas sequenciadas nas corridas de validação.

Tabela 2: Corridas realizadas e a respetiva distribuição das amostras ao longo destas. Amostras negativas incluídas nas corridas de validação também tiveram seus resultados finais utilizados nas análises de ganho diagnóstico.

Corrida	Quantidade de Amostras (Pos/Neg)	Replicata NA12878	Validação	Ganho Diagnóstico
1	5 (2/3)		x	x
2	5 (1/4)		x	x
3	5 (2/3)		x	x
4	5 (5/0)		x	
5	5 (3/2)		x	x
6	5 (1/4)		x	x
7	5 (3/2)		x	x
8	5 (0/5)			x
9	5 (0/5)			x
10	5 (0/5)			x
11	5 (2/3)		x	x
12	5 (0/5)			x
13	5 (0/5)			x
14	5 (0/5)			x
15	4 (0/4)			x
16	5 (0/4)	NA12878_1	x	x
17	5 (0/4)	NA12878_2	x	x
18	5 (0/4)	NA12878_3	x	x
19	5 (0/5)			x
20	5 (2/0)	NA12878_1.1, _2.1 e _3.1	x	

3.5 Processo de enriquecimento

O protocolo de enriquecimento das amostras foi realizado de acordo com as instruções do fabricante. A Figura 3 ilustra todas as etapas pelas quais as amostras são processadas. Antes do enriquecimento, as amostras eram quantificadas com a utilização do fluorómetro Qubit 2.0 (Life Technologies) e diluídas a uma concentração de 5ng/μL, em 50μL. Após um processo de fragmentação enzimática, as amostras eram misturadas às sondas e a *barcodes*, e então colocadas num termociclador por 16 horas a 54°C. Ao final dessa etapa, as sondas (agora já ligadas às regiões de interesse) eram

capturadas com o uso de grânulos ferromagnéticos (*beads*) revestidos com estreptavidina, de tal maneira que eram separadas das demais regiões gênicas que não seriam analisadas. Após essa etapa, uma eluição com NaOH a 50mM separava as regiões do DNA capturado das sondas, e o material genético proveniente dessa separação era amplificado em uma PCR de 20 ciclos. Após amplificação, o material era purificado e eluído com Tris HCL 10mM (pH = 8,0) para armazenamento. A quantificação das amostras e a avaliação de tamanho dos fragmentos capturados foram realizadas com o equipamento Bioanalyzer (Agilent Technologies).

3.6 Preparo de *template* e sequenciamento

O preparo de *template* foi realizado com o equipamento Ion One Touch System e o kit Ion OT2 200 Template Kit (Life Technologies). Após a quantificação, as amostras eram diluídas a uma concentração de 10pM e cinco amostras diferentes eram misturadas em um *pool* equimolar e 25µL eram utilizados para preparo do *template*. Após essa etapa, o *template* era enriquecido no equipamento Ion One Touch ES (Life Technologies) para evitar a entrada de *beads* sem DNA no sequenciamento.

As amostras foram então sequenciadas no equipamento Ion Torrent PGM™ (Life Technologies), e todas as corridas foram realizadas com o kit Ion PGM 200 Sequencing Kit v2. Cada corrida consistia no sequenciamento de cinco amostras de DNA com a utilização do chip Ion 318, o qual apresenta um desempenho esperado de 1 Gigabase (Gb) de resultado.

Figura 3: Esquema ilustrando os processos de preparo, sequenciamento e análises de bioinformática aos quais as amostras eram submetidas. Os quadrados com borda serrilhada indicam etapas envolvidas no processo de validação do *pipeline*.

3.7 Análises de Bioinformática

Ao final de cada corrida, as sequências geradas eram processadas para a remoção de sequências de adaptadores e geração dos arquivos FASTQ. Após a geração desses arquivos, os mesmos eram importados para o software CLC Genomics Workbench 7 (CLCbio, QIAGEN) no qual um *pipeline* foi montado para as análises de bioinformática. As principais etapas eram:

- Uma etapa inicial de tratamento das sequências no qual eram removidos 5pb do final 3' de cada sequência gerada, para que não fossem

mapeadas bases com baixa acurácia. Sequências curtas e com valores de qualidade Phred abaixo de Q20 também eram removidas.

- Mapeamento das sequências contra o genoma humano referência (hg19) com valores padrão de penalização para *mismatches* (2), InDels (3) e fração mínima de similaridade entre a sequência alinhada e a referência (80%). Um relatório de mapeamento e um arquivo de visualização BAM eram gerados após essa etapa.
- Geração de estatísticas para as regiões alvo para identificação de regiões com baixa cobertura de sequenciamento. As regiões alvo eram definidas como exons mais 10pb de regiões intrônicas flangeadoras. Um relatório de cobertura era gerado ao final do processo.
- Chamada de variantes baseada em qualidade de sequenciamento, o qual utilizava a qualidade de sequenciamento da vizinhança (raio de 5pb da possível base alterada) como um critério para chamadas positivas. A presença bidirecional de variantes era requisitada com uma concordância mínima de 5% entre sequências *Forward* e *Reverse*. Os limiares de cobertura mínima de sequenciamento e frequência de alelo variante (FAV) no processo de validação, foi testada com as seguintes combinações: cobertura mínima de 30 vezes (30x) com FAV de 35%, 25% e 20%, e cobertura mínima de 10x com FAV de 35% e 20%. Os valores de cobertura mínima e FAV foram estipulados de acordo com o aparecimento em outros estudos^{40,29,41,28} bem como a requisição da presença das variantes nos sentidos F e R como uma forma de se evitar falsas descobertas⁴².

- Processo de anotação dividido em 3 etapas: anotação com informações dos genes (nomes e códigos referência); anotação com números dos exons nos quais as mutações se encontravam; anotação com os dados do dbSNP138 para verificação de alterações já descritas.
- Predição de alterações não-sinônimas e de sítio de *splicing*. Uma tabela contendo todas as variantes encontradas em cada amostra era gerada ao final desse processo no formato XLS.

O *pipeline* montado apresentou um tempo de análise de 30 minutos por amostra (2h30min por corrida). O software foi instalado em um servidor com 48GB de memória RAM com dois processadores de seis núcleos.

3.8 Processo de validação e avaliação analítica

As amostras da casuística de CH do InCor, envolvidas no processo de validação, tiveram seus resultados finais filtrados apenas para as regiões dos 3 genes sequenciados posteriormente (*MYH7*, *MYBPC3* e *TNNT2*), de tal maneira que apenas as alterações dentro das regiões previamente sequenciadas por Sanger foram levadas em consideração. A sensibilidade foi calculada como Verdadeiros positivos (VP)/ Total de VP, sendo que VP são cada uma das alterações detectadas pelo método Sanger, considerado padrão-ouro para sequenciamento.

A lista *benchmark* de variantes da amostra NA12878 foi baixada num formato denominado *Variant Call Format* (VCF) através de um endereço de servidor FTP³⁹ do *National Center for Biotechnology Information* (NCBI, EUA). Uma lista *benchmark* foi obtida com variantes hetero e homozigotas de alta confiabilidade, bem como um arquivo contendo toda as regiões com alta

qualidade de sequenciamento da amostra. Dessa forma, todas as variantes detectadas no sequenciamento da amostra NA12878, dentro do *pipeline* proposto, puderam ser comparadas com a lista *benchmark*, gerando um resultado de acurácia.

A comparação entre os dados gerados no *pipeline* e a lista *benchmark* foi feita através do programa *VCFcomparator*⁴³. No momento da comparação, um dos requisitos para que uma variante fosse considerada verdadeira foi a concordância genotípica, ou seja, apenas variantes detectadas no genótipo correto (hetero ou homozigoto) foram consideradas como VP. Aquelas detectadas com o genótipo incorreto foram consideradas “não concordantes” (NC). Variantes não detectadas foram consideradas Falsos Negativos (FN) e detectadas apenas no *pipeline* testado (ou seja, ausentes da lista *benchmark*) foram consideradas Falsos Positivos (FP).

Dessa forma, os parâmetros analíticos foram calculados como: Sensibilidade = $VP / \text{Total de variantes } benchmark$; Especificidade = $\text{Verdadeiros Negativos (VN)} / \text{VN} + \text{FP}$; Valor Preditivo Positivo (VPP) = $VP / \text{VP} + \text{NC} + \text{FP}$; Taxa de Falsas Descobertas (TFD) = $1 - \text{VPP}$. A Reprodutibilidade foi definida como a concordância de VP entre as triplicatas da amostra NA12878, tanto para as análises intra como para inter-ensaio.

3.9 Análise de ganho diagnóstico

Todas as amostras negativas sequenciadas e analisadas no processo descrito foram submetidas à análise de ganho diagnóstico, na tentativa de se encontrar alguma alteração que pudesse ser indicada como possivelmente causadora do fenótipo de hipertrofia cardíaca. Dado que essas amostras

tiveram resultado negativo no sequenciamento dos três genes mais associados à CH, o propósito dessa etapa foi quantificar o aumento de casos descritos como positivos, uma vez que agora foram analisados 74 genes relacionados com cardiomiopatias hereditárias.

Inicialmente, alterações preditas como causadoras de modificação na sequência das proteínas eram separadas das demais (Figura 4). Dentre essas, constavam alterações causadoras de troca de aminoácido (*missense*), mudança no quadro de leitura do gene (*frameshift*), geradoras de códon de parada (*nonsense*) e possivelmente modificadoras de sítio de *splicing*. Após essa primeira filtragem, as alterações *missense* remanescentes eram submetidas à três algoritmos de predição de impacto funcional: *Protein Variation Effect Analyzer* (PROVEAN)⁴⁴, *Sort Intolerant from Tolerant* (SIFT)⁴⁵ e *Polymorphism Phenotyping v2* (PolyPhen-2)⁴⁶, os quais indicavam o grau de impacto da substituição do aminoácido original pelo variante através da análise de conservação evolutiva da proteína analisada. Eram também computados os valores obtidos através do programa *Combined Annotation Dependent Depletion* (CADD) para todos os tipo de alteração⁴⁷, sendo que são consideradas deletérias alterações com valor acima de 15 e, quanto maior o valor, maior a probabilidade de malignidade.

A frequência populacional das alterações era consultada através da anotação com os dados do *Single Nucleotide Polymorphism Database* (dbSNP)⁴⁸, e também através dos bancos de dados dos projetos *Exome Sequencing Project* (ESP)³⁴ e *Exome Aggregation Consortium* (ExAC)⁴⁹. Os bancos de dados *Human Genome Mutation Database* (HGMD)⁵⁰ e ClinVar⁵¹

foram consultados para verificação de associação prévia de variantes com a doença estudada. A princípio eram analisadas as alterações em heterozigose e, caso não houvesse nenhuma alteração potencialmente patogênica, checava-se então as alterações em homozigose.

Assim, após todas essas análises, eram consideradas potencialmente patogênicas as alterações que respeitassem os seguintes critérios:

- Alterações em genes previamente associados à CH;
- Alterações com frequência populacional do alelo menor abaixo de 1% (ou MAF, do inglês, *Minor Allele Frequency*) ou não descritas;
- Impacto funcional considerado deletério ou danoso para a proteína em pelos dois algoritmos testados;
- Associação prévia descrita na literatura.

Figura 4: Algoritmo de análise para alterações encontradas nos resultados de SNG.

4. RESULTADOS

4.1 Desempenho Geral do Sequenciamento

O sequenciamento na plataforma Ion Torrent PGM apresentou um rendimento médio de $165,9 \pm 13,1$ Mb, com um valor médio de $146,9 \pm 11,54$ Mb acima de $\text{PhredQ} \geq 20$, por amostra (Tabela 3). Estatísticas de alinhamento mostraram que em média $93,83 \pm 0,68\%$ das sequências geradas foram mapeadas contra a referência sendo que a especificidade média para regiões-alvo foi de $93,54 \pm 0,28\%$. A taxa de policlonalidade apresentou valores mínimos e máximos de 7% e 26%, respectivamente.

Tabela 3: Desempenho do sequenciamento na plataforma Ion Torrent PGM™

Corridas	Desempenho (Mb)	Média de Mb por amostra	Média de Q $\geq$ 20 Mb por amostra	Policlonalidade
1	834	173,0	144,5	17%
2	1000	212,4	186,8	19%
3	885	173,3	156,6	22%
4	1100	235,4	211,8	19%
5	961	191,3	168,2	21%
6	824	163,8	146,7	20%
7	937	181,3	155,7	7%
8	779	154,5	136,2	25%
9	977	194,4	174,8	19%
10	985	195,8	176,4	13%
11	794	157,5	137,2	23%
12	716	141,5	122,9	7%
13	823	163,4	140,3	26%
14	862	171,7	148,4	22%
15	883	150,0	132,9	26%
16	608	120,4	111,1	19%
17	578	114,1	104,6	19%
18	871	172,1	157,3	26%
19	540	106,7	98,1	19%
20	704	140,0	123,9	15%
Média$\pm$IC95%	833,05$\pm$64,7Mb	165,9$\pm$13,1Mb	146,9$\pm$11,54Mb	19$\pm$2,39%

4.2 Validação analítica

4.2.1 Cobertura de Regiões Alvo

Para o cálculo de cobertura foram consideradas as 52 amostras da casuística de CH, as 6 replicatas da amostra NA12878 e as repetições das amostras P11 e P13, resultando em 60 amostras. A cobertura média das amostras utilizadas no processo de validação foi de $250 \pm 23,94x$. Como descrito previamente²⁸, o Fator de Enriquecimento foi calculado e apresentou valor médio foi de $7930 \pm 56,27$. Com o propósito de se visualizar a distribuição dos valores de cobertura, um histograma foi montado utilizando os valores de cobertura média de cada uma das 1754 regiões alvo (Figura 5). Sessenta e três regiões alvo (3,5%) se apresentaram dentro do intervalo entre 0 e 30x de cobertura, enquanto apenas 28 regiões (1,6%) se apresentaram dentro do intervalo entre 0 e 10x.

Figura 5: Distribuição dos valores de cobertura média das 1754 regiões alvo. A barra preta indica as regiões com cobertura média abaixo de 10x.

A cobertura de todas as bases interrogadas no painel foi analisada dentro de um espectro de 7 pontos de cobertura, e os resultados estão apresentados na Tabela 4.

Tabela 4: Porcentagem de cobertura do painel analisada em 7 pontos diferentes de cobertura.

Cobertura	Quantidade de bases	IC 95%
1x	97,7%	97,57 - 97,81
5x	96,5%	96,27 - 96,75
10x	95,2%	94,77 - 95,57
20x	92,5%	91,75 - 93,27
25x	91,1%	90,2 - 92,08
50x	84,2%	82,43 - 85,91
100x	70,4%	67,52 - 73,32

A distribuição de cobertura das amostras envolvidas no processo de validação foi analisada através de um gráfico boxplot, no qual é possível analisar a variabilidade inter-amostral (Figura 6).

Figura 6: Distribuição de cobertura das amostras envolvidas no processo de validação. Os grupos com cores intercaladas indicam quais amostras foram processadas dentro da mesma corrida.

4.2.2 Sensibilidade e Especificidade

Da casuística de CH, foram analisadas as 285 alterações presentes nas 52 amostras sequenciadas. Desse total, 20 alterações eram patogênicas (14 mutações pontuais, 3 deleções e 3 variantes em sítio de *splicing*) e 265 eram polimorfismos. Exceto pelas alterações em sítio de *splicing*, todas as outras mutações e SNVs ocorriam dentro de regiões codificantes.

Inicialmente, foi fixado um parâmetro de cobertura mínima de 30x para a chamada de variantes e, a partir desse valor, a sensibilidade foi testada com três diferentes valores de FAV: 35%, 25% e 20%. Os valores obtidos foram de 92,3% (263/285) com a FAV em 35% e 94,0% (268/285) com 25% e 20%. Em seguida, o parâmetro de cobertura mínima foi modificado para 10x e testado com FAVs de 35% e 20%, sendo que os resultados obtidos foram de 94,7% (270/285) e 96,5% (275/285), respectivamente (Tabela 5).

Tabela 5: Valores de sensibilidade e especificidade da chamada de variantes nos genes MYH7, MYBPC3 e TNNT2, com as amostras da casuística de CH previamente processadas no Painel CH-Sanger.

Cobertura Mínima				30x		10x	
FAV	35%	25%	20%	35%	20%		
VP	263	268	268	270	275		
FP	0	0	0	0	0		
Sensibilidade	92,3%	94%	94%	94,7%	96,5%		
IC 95%	89,55 – 96,45	91,41 – 97,48	91,41 – 97,48	92,94 – 97,75	94,7 – 98,92		
Especificidade	100%	100%	100%	100%	100%		
IC 95%	99,0 - 100	99,0 - 100	99,0 - 100	99,0 - 100	99,0 – 100		

De todas as alterações patogênicas, 18 foram confirmadas (90%, sendo 16 mutações pontuais e duas deleções) com uma boa cobertura e dentro do genótipo correto (Tabela 6). Nenhum FP foi identificado dentro das regiões

posteriormente analisada pelo sequenciamento Sanger, indicando uma especificidade de 100% para essas regiões.

Tabela 6: Mutações patogênicas presentes nas amostras da casuística de CH e suas respectivas confirmações.

Amostras	Mutação Patogênica	Confirmada	Cobertura (FAV%)
P1	MYH7 - p.Ser842Gly	Sim	249x (37,7%)
P2	MYBPC3 - c.3330+2T>C	Sim	396x (50,2%)
P3	MYH7 - p.Ile263Thr	Sim	895x (44,5%)
P4	MYH7 - p.Ile702Val	Sim	242x (46,7%)
P5	MYH7 - p.Phe764Tyr	Sim	127x (54,3%)
P6	MYH7 - p.Ile702Val	Sim	329x (50%)
P7	MYH7 - p.Phe252Cys	Sim	340x (41,1%)
P8	MYH7 - p.Arg249Gln	Sim	599x (45%)
P9	MYH7 - p.Arg858Cys	Sim	139x (43,1%)
P10	MYH7 - p.Val320Met	Sim	227x (48,4%)
P11	MYH7 - p.Met493Ile	Sim	114x (53,5%)
P12	MYBPC3 - c.2905+1G>A	Sim	155x (52,2%)
P13	MYH7 - p.Lys1459Asn	Sim	262x (40%)
P14	MYBPC3 - c.3628-1G>A	Sim	136x (43,4%)
P15	MYBPC3 - Phe305fs	Não	88x (80%)*
P16	MYH7 - p.Val320Met	Sim	126x (47,6%)
P17	MYBPC3 - p.Glu619Lys / MYBPC3 - p.Leu1221fs	Não/Sim	9x (89%) / 203x (48,2%)
P18	MYH7 - p.Ala797Thr	Sim	334x (48,5%)
P19	MYBPC3 - p.Asn850fs	Sim	37x (48%)

*variante sem representação bidirecional

A análise completa do painel foi realizada com os dados gerados através do sequenciamento da amostra NA12878. Ao todo, 311728pb foram passíveis de comparação entre os dados *benchmark* da amostra e as regiões de interesse do painel, e os mesmos parâmetros de chamada de variantes, com as mesmas combinações, foram testado nesses dados. As Tabelas 7 e 8 mostram os resultados das validações inter e intra-ensaio, utilizando o programa *VcfComparator*.

Assim como observado com as amostras da casuística de CH, a sensibilidade aumentou com a seleção de parâmetros menos estridentes, com a combinação de 10x de cobertura e 20% de FAV sendo o ponto de sensibilidade máxima. O valor médio de sensibilidade nas validações inter e intra-ensaio foram de 92,7% e 91%, respectivamente, sendo que a sensibilidade máxima observada foi de 94,7% para a replicata NA12878_2 na corrida 10. A validação apresentou baixa ocorrência de variantes NC, FP, e valores de especificidade uniformes ao longo de todo o processo, com o máximo valor observado de 0,9999967.

Analisando a sensibilidade de forma estratificada pelo tipo de variante em ambos os materiais, 560 SNVs (282 da casuística de CH, 132 da validação inter-ensaio, 132 da validação intra-ensaio, e 14 da repetição das amostras P11.1 e P13.1) e 7 InDels (3 da casuística de CH, 2 da validação inter-ensaio e 2 da validação intra-ensaio) deveriam ser confirmados. De todos os SNVs foram confirmados 273 da casuística de CH, 127 da replicata NA12878_2 na validação inter-ensaio, 126 da replicata NA12878_2.1 na validação intra-ensaio e 14 das amostras P11.1 e P13.1. Assim, a sensibilidade máxima para SNVs foi de 96,4% (540/560). Se adicionarmos a essa conta os casos NCs (como se fossem chamadas verdadeiras), seriam adicionados 2 SNVs ao valor final (providos da amostra NA12878_2.1) aumentando a sensibilidade para 96,7% (542/560). Das 7 InDels a serem confirmadas, apenas duas, ambas com 1pb de tamanho, foram confirmadas. Quatro InDels (uma com 2pb da amostra P15, uma inserção de 7pb e duas deleções de 5pb da amostra NA12878) não foram confirmadas devido a problemas de cobertura de sequenciamento e uma

inserção de 7pb das amostras NA12878_1.1, _2.1 e _3.1 foi chamada incorretamente não sendo considerada um VP. Dessa foram, a sensibilidade para deleções foi de 28,5% (2/7).

4.2.3 VPP, TFD e Reprodutibilidade

O VPP foi calculado para cada triplicata da NA12878 em cada etapa da chamada de variantes, tanto para as análises inter quanto intra-ensaio. Esses valores permaneceram acima da 0,959 em todas os cenários de chamada de variantes, com um valor médio de 0,977. Os valores de TFD (definida como $1 - VPP$) apresentaram uma média de 0,021 e o valor máximo observado de 0,41.

A reprodutibilidade foi avaliada como a concordância entre os VP confirmados em cada uma das replicatas em cada um dos cenários de chamada de variantes. A avaliação com a amostra NA12878 resultou em um valor de reprodutibilidade inter e intra-ensaio de 89,5% e 87,3%, respectivamente (Figura 7). A diminuição progressiva dos valores de cobertura e FAV na chamada de variantes influenciou diretamente os valores reprodutibilidade, sendo que o valor máximo foi observado no cenário de 10x de cobertura mínima e 20% de FAV. As amostras P11.1 e P13.1 apresentaram 100% de reprodutibilidade, com as 14 variantes sendo identificadas em ambas as corridas.

Figura 7: Análise de reprodutibilidade entre as replicatas da amostra NA12878, nas perspectivas inter-ensaio (Painel A) e intra-ensaio (Painel B). Ao todo 134 variantes constavam na tabela *benchmark* para serem confirmadas.

Tabela 7: Resultados da validação analítica inter-ensaio realizada com a amostra NA12878. Abreviações: Não concordante (NC); Valor Preditivo Positivo (VPP); Taxa de Falsas Descobertas (TFD); Reprodutibilidade (Rep) e Intervalo de Confiança (IC).

	30x							10x						
	35%							35%						
	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep
NA12878_1	112 (83,5%)	2	1	0,9999967	0,973	0,026	108 (80,5%)	120 (89,5%)	2	1	0,9999967	0,975	0,024	113 (84,3%)
NA12878_2	117 (87,3%)	0	1	0,9999967	0,991	0,008		126 (94%)	0	1	0,9999967	0,992	0,0078	
NA12878_3	114 (85%)	2	1	0,9999967	0,974	0,025		118 (88%)	3	1	0,9999967	0,967	0,032	
IC 95%	83,1 - 87,43			0,99998 - 1	0,97 - 0,99			86,97 - 94,03			0,99998 - 1	0,96 - 0,99		
	25%													
	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep							
NA12878_1	115 (85,5%)	1	1	0,9999967	0,982	0,017	111 (82,8%)							
NA12878_2	117 (87,3%)	0	1	0,9999967	0,991	0,008								
NA12878_3	115 (85,5%)	1	1	0,9999967	0,982	0,017								
IC 95%	84,92 - 87,28			0,99998 - 1	0,98 - 0,99									
	20%							20%						
	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep
NA12878_1	115 (85,5%)	1	2	0,9999935	0,974	0,017	112 (83,5%)	124 (92,5%)	1	2	0,9999935	0,976	0,023	120 (89,5%)
NA12878_2	117 (87,3%)	0	1	0,9999967	0,991	0,008		127 (94,7%)	0	1	0,9999967	0,992	0,0078	
NA12878_3	116 (86,5%)	0	1	0,9999967	0,991	0,0085		122 (91%)	0	1	0,9999967	0,991	0,008	
IC 95%	85,41 -87,45			0,99998 - 1	0,97 - 1			90,63 - 94,84			0,99998 - 1	0,98 - 1		

Tabela 8: Resultados da validação analítica intra-ensaio realizada com a amostra NA12878. Abreviações: Não concordante (NC); Valor Preditivo Positivo (VPP); Taxa de Falsas Descobertas (TFD); Reprodutibilidade (Rep) e Intervalo de Confiança (IC).

30x								10x						
35%								35%						
	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep
NA12878_1.1	109 (81,3%)	1	1	0,9999967	0,981	0,019	105 (78,3%)	118 (88%)	3	2	0,9999935	0,959	0,041	113 (84,3%)
NA12878_2.1	123 (91,7%)	1	1	0,9999967	0,984	0,016		126 (94%)	2	1	0,9999967	0,976	0,024	
NA12878_3.1	111 (82,8%)	2	1	0,9999967	0,973	0,027		118 (88%)	4	1	0,9999967	0,959	0,041	
IC 95%	78,91 - 91,6			0,99998 - 1	0,97 - 0,99	0,01 - 0,03		86,08 - 93,92			0,99998 - 1	0,95 - 0,98	0,02 - 0,05	
25%														
	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep							
NA12878_1.1	112 (83,5%)	0	2	0,9999935	0,982	0,018	108 (80,5%)							
NA12878_2.1	123 (91,7%)	1	1	0,9999967	0,984	0,016								
NA12878_3.1	111 (82,8%)	2	1	0,9999967	0,973	0,027								
IC 95%	80,4 - 91,6			0,99998 - 1	0,97 - 0,99	0,01 - 0,03								
20%								20%						
	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep	Sensibilidade	NC	FP	Especificidade	VPP	TFD	Rep
NA12878_1.1	112 (83,5%)	0	2	0,9999935	0,982	0,018	108 (80,5%)	121 (90,2%)	2	3	0,9999904	0,96	0,04	117 (87,3%)
NA12878_2.1	123 (91,7%)	1	2	0,9999935	0,976	0,024		126 (94%)	2	2	0,9999904	0,969	0,031	
NA12878_3.1	111 (82,8%)	2	1	0,9999967	0,973	0,027		119 (88,8%)	3	2	0,9999935	0,959	0,041	
IC 95%	80,4 - 91,6			0,99998 - 1	0,97 - 0,98	0,02 - 0,03		87,96 - 94,04			0,99998 - 1	0,96 - 0,97	0,03 - 0,04	

4.2.4 Investigação de variantes não encontradas e discrepantes

Uma vez determinado que o cenário de 10x de cobertura mínima e 20% de FAV foi o ponto de maior sensibilidade da chamada de variantes, todas as variantes não encontradas e NCs nesse cenário foram investigadas visualmente com o uso do programa *Integrative Genomics Viewer (IGV)*⁵², na tentativa de se encontrar uma possível razão para a ausência dessas no resultado final.

De todas as alterações previstas na casuística de CH, 10 permaneceram ausentes dos resultados de validação, sendo que duas eram mutações patogênicas e 8 eram polimorfismos (Tabela 9). O principal motivo da perda de variantes nessa etapa da validação foi a ausência de representatividade bidirecional das variantes (7 variantes), seguido de baixa cobertura devido ao alto conteúdo GC da região (2 variantes) e um caso de total ausência do alelo variante na posição.

Tabela 9: Variantes perdidas no processo de validação com as amostras da casuística de CH.

Amostra	Variante Perdida	Gene	Corrida	Motivo
P2	rs35078470	MYBPC3	1	79.6% GC (9x cob)
P8	rs3729953	MYBPC3	4	Ausência de AV (96x cov)
P15	F305fs	MYPBC3	7	Sem cobertura bidirecional
P17	E619K	MYBPC3	7	73.7% GC (9x cob)
N2	rs2069540	MYH7	1	Sem cobertura bidirecional
N4	rs3729547	TNNT2	2	Sem cobertura bidirecional
N5	rs3729547	TNNT2	2	Sem cobertura bidirecional
N6	rs3729547	TNNT2	2	Sem cobertura bidirecional
N7	rs2069540/rs3729547	MYH7/TNNT2	2	Sem cobertura bidirecional

Em relação às análises feitas com a amostra NA12878, a avaliação intra-ensaio não detectou 17 alterações e apresentou 7 variantes discrepantes, enquanto que na avaliação inter-ensaio esses números se apresentaram diminuídos, com 14 variantes perdidas e 4 discrepantes (Tabelas 7 e 8). As causas da perda de variantes foram praticamente as mesmas observadas nas amostras da casuística de CH (Tabelas 10 e 11). Baixas taxas de cobertura associadas a regiões com alto ou baixo conteúdo GC foram frequentemente observadas. Ausência de representação bidirecional também foi observada, porém em menor número.

Tabela 10: Análise das variantes não encontradas e discrepantes no experimento inter-ensaio com a amostra NA12878.

Variantes referência não encontradas (10x/20%)					
Crom	Pos	Ref/Var	Replicata	Gene	Motivo
1	201331068	A/G	1	<i>TNNT2</i>	Genótipo Não-concordante
1	201341175	CAGAAG/C	1,2 e 3	<i>TNNT2</i>	Cobertura < 10x
1	237957161	A/G	1,2 e 3	<i>RYSR2</i>	Cobertura < 10x (36.9% GC)
2	220283259	A/G	3	<i>DES</i>	Cobertura < 10x (72.2% GC)
2	220283277	T/C	3	<i>DES</i>	Cobertura < 10x (72.2% GC)
2	179614952	A/G	1 e 3	<i>TTN</i>	Cobertura < 10x
6	7556063	C/T	3	<i>DSP</i>	Sem cobertura bidirecional
15	63351840	C/A	1,2 e 3	<i>TPM1</i>	Cobertura < 10x
17	78081526	A/AGCAGCGG	1,2 e 3	<i>GAA</i>	Cobertura < 10x (71.6% GC)
18	28611139	A/C	1,2 e 3	<i>DSC3</i>	Cobertura < 10x (25% GC)
18	29122799	G/A	1,2 e 3	<i>DSG2</i>	Sem cobertura bidirecional
20	30414621	C/T	1 e 2	<i>MYLK2</i>	Sem cobertura bidirecional
20	42744587	G/C	1 e 3	<i>JPH2</i>	Cobertura < 10x (75.1% GC)
X	119590533	T/A	3	<i>LAMP2</i>	Cobertura < 10x (32.3% GC)

Variantes discrepantes nos resultados experimentais (10x/20%)					
Crom	Pos	Ref/Var	Replicata	Gene	Motivo
1	201331068	A/G	1	<i>TNNT2</i>	Genótipo Não-concordante
2	179568916	T/G	2 e 3	<i>TTN</i>	Falso Positivo
6	7580724	A/G	1	<i>DSP</i>	Falso Positivo
10	121436245	A/AC	1	<i>BAG3</i>	Falso Positivo

Tabela 11: Análise das variantes não encontradas e discrepantes no experimento intra-ensaio com a amostra NA12878.

Variantes referência não encontradas (10x/20%)					
Crom	Pos	Ref/Var	Replicata	Gene	Motivo
1	201331068	A/G	1,2 e 3	<i>TNNT2</i>	Cobertura < 10x
1	201341175	CAGAAG/C	1,2 e 3	<i>TNNT2</i>	Cobertura < 10x
1	237957161	A/G	1,2 e 3	<i>RYS2</i>	Cobertura < 10x (36.9% GC)
2	179497133	C/T	1 e 3	<i>TTN</i>	Cobertura < 10x (31.9% GC)
2	179614952	A/G	1 e 3	<i>TTN</i>	Cobertura < 10x
6	152647681	A/T	3	<i>SYNE1</i>	Cobertura < 10x (36.6% GC)
7	128484816	A/G	3	<i>FLNC</i>	Sem cobertura bidirecional
14	64491695	T/C	1	<i>SYNE2</i>	Cobertura < 10x (33.4% GC)
14	64496749	C/T	1	<i>SYNE2</i>	Cobertura < 10x
15	63351840	C/A	1,2 e 3	<i>TPM1</i>	Cobertura < 10x
17	78081526	A/AGCAGCGG	1,2 e 3	<i>GAA</i>	Deteção incorreta
18	28611139	A/C	1,2 e 3	<i>DSC3</i>	Cobertura < 10x (25% GC)
18	29104698	C/T	1 e 3	<i>DSG2</i>	Cobertura < 10x (37.8% GC)
18	29122799	G/A	1,2 e 3	<i>DSG2</i>	Sem cobertura bidirecional
20	30414621	C/T	3	<i>MYLK2</i>	Genótipo Não-concordante
20	42744587	G/C	1,2 e 3	<i>JPH2</i>	Genótipo Não-concordante
X	119590533	T/A	3	<i>LAMP2</i>	Cobertura < 10x (32.3% GC)

Variantes discrepantes nos resultados experimentais (10x/20%)					
Crom	Pos	Ref/Var	Replicata	Gene	Motivo
6	7572262	A/G	1,2 e 3	<i>DSP</i>	Falso Positivo
6	7584618	G/T	1	<i>DSP</i>	Falso Positivo
6	152545709	C/T	3	<i>SYNE1</i>	Falso Positivo
12	98941542	C/CT	1 e 2	<i>TMPO</i>	Falso Positivo
17	78081526	A/AGCAGC	1,2 e 3	<i>GAA</i>	Deteção incorreta
20	42744587	G/C	1,2 e 3	<i>JPH2</i>	Genótipo Não-concordante
20	30414621	C/T	3	<i>MYLK2</i>	Genótipo Não-concordante

Duas questões podem levar a ocorrência de alterações discrepantes das apresentadas na planilha *benchmark*. A primeira é o genótipo não concordante, o que não permite que a variante seja reconhecida como um VP, apesar de ser uma alteração genuína. A segunda é o aparecimento de FPs, as quais são de fato alterações não presentes na amostra NA12878, e podem indicar vieses do método. Ambos os eventos foram analisados e se mostraram de baixa ocorrência (Tabelas 7 e 8).

No cenário de maior sensibilidade, o valor máximo observado foi de 3 tanto para NCs (NA12878_3.1) quanto para FPs (NA12878_1.1). Das sete regiões implicadas na ocorrência de FPs, quatro eram regiões de homopolímero.

4.3 Ganho diagnóstico

As 72 amostras negativas foram analisadas dentro do critério descrito na seção “Materiais e Métodos”. Do total, 35 amostras (49%) foram reclassificadas como positivas, 22 amostras como inconclusivas (30%) e 15 amostras (21%) permaneceram negativas (Figura 8).

Figura 8: Proporção de reclassificação das 72 amostras negativas sequenciadas.

As Tabelas 12 e 13 apresentam as alterações encontradas nas amostras consideradas positivas e inconclusivas, respectivamente, bem como todas as informações utilizadas no processo de decisão.

Tabela 12: Alterações encontradas nas amostras reclassificadas como positivas. Abreviações: Cardiomiopatia hipertrófica (CH); Amiloidose (Amil.); Síndrome de Noonan (Noonan); PROVEAN (Pr); SIFT (S); PolyPhen-2 (P2).

Amostra	Gene	Alteração	Deletéria			CADD Score	HGMD Assoc.	ClinVar	MAF		
			Pr	S	P2				ExAC	EVS	dbSNP
N1	TRIM63	p.Cys145Tyr	x	x	x	32	-	-	-	-	0,001
N5 [‡]	MYL2	p.Gly42Asp	x	x	x	29,9	-	-	-	-	-
N7	FLNC	p.Tyr1230Cys	x	x	x	27,4	-	-	-	-	-
	TRIM63	p.Cys142Tyr	x	x	x	32	-	-	3,29E-05	1,53E-04	0,0002
N14	NEXN	p.Ser596Arg		x	x	12,9	-	-	-	1,67E-04	0,0002
N16	TNNI3	p.Arg145Trp	x	x	x	20,3	CH	Patogênica	8,29E-06	-	-
N17	FLNC	p.Ser1624Leu	x	x	x	22,1	-	-	-	-	-
N18	PRKAG2	p.Gly100Ser			x	22	Amil.	Prov. Patog.	0,0081	8,46E-04	0,0146
N19	ACTC1	p.Glu101Lys	x	x	x	22,4	CH	Patogênica	-	-	-
N20	VCL	p.Thr263Ser		x	x	23,4	-	-	6,61E-05	2,30E-04	0,0006
N21	TPM1	p.Arg21Leu	x	x	x	22,5	-	-	0,0001104	-	-
	ACTN2	p.Ser369Leu	x	x	x	34	-	-	6,79E-05	-	-
N26	MYL3	p.Met173Thr	x	x		25,4	CH	-	-	-	-
N27	FLNC	p.Thr1681Arg		x	x	19,3	-	-	0,0005719	0,0014	0,0026
N28	TTR	p.Val142Ile		x	x	21,7	Amil.	Patogênica	0,001376	0,005	0,0056
N31	TRIM63	p.Gln247*	-	-	-	38	CH	-	0,000486	4,61E-04	0,0002
N34 [‡]	MYL2	p.Gly42Asp	x	x	x	29,9	-	-	-	-	-
N36	TTR	p.Val50Met		x	x	25,1	Amil.	Patogênica	0,0001483	-	-
N37	MYH6	p.Asp377Glu	x	x		4,9	-	Prov. Benigna	7,17E-04	0,0022	0,0022
N39	TNNT2	p.Trp294*	-	-	-	45	CH	-	-	-	-
N40	TTN	p.Glu9502del [¥]	-	-	-	22,2	-	-	-	-	-
	MYBPC3	p.Gly507Arg	x	x	x	33	CH	Benigna	0,00067	0,0013	0,0026
N41	MYBPC3	p.Gly507Arg	x	x	x	33	CH	Benigna	0,00067	0,0013	0,0026
	TNNT2	p.Trp294*	-	-	-	45	CH	-	-	-	-
N43	VCL	p.Asp511Tyr	x	x	x	34	-	-	-	-	-
N46	MYH7	p.Glu1120Lys	x	x	x	22,3	-	-	-	-	-
N49	RAF1	p.Pro261Ala	x	x	x	14,9	Noonan	Patogênica	-	-	-
N50	TNNI3	p.Ser199Asn		x	x	22,4	CH	-	-	-	-
N51	MYL3	p.Val156Leu	x		x	22,5	CH	Prov. Patog.	-	-	-
	TPM1	p.Arg21Leu	x	x	x	22,5	-	-	0,0001104	-	-
N52	FLNC	p.Ala2041Thr	x	x	x	22,4	-	-	-	-	-
N53	TNNI3	p.Leu198Val			x	19,7	CH	-	-	-	-
	FLNC	p.Asp693Ala	x	x	x	29,4	-	-	0,003702	0,004	0,002
N54	MAP2K2	p.Gly214Arg	x	x	x	32	-	-	1,57E-05	-	-
N55 [‡]	TRIM63	p.Gln247*	-	-	-	38	CH	-	0,000486	4,61E-04	0,0002
N60	FLNC	p.Tyr1519Cys	x	x	x	17,65	-	-	-	-	-
N67	MYL3	p.Asn145Thr	x	x	x	22,5	-	-	-	-	-
	TPM1	p.Arg21Leu	x	x	x	22,5	-	-	0,0001104	-	-
N68	ANK2	p.Glu3062Gly	x	x	x	28	-	-	-	-	-
N69	MYH6	p.Asp377Glu	x	x		4,9	-	Prov. Benigna	7,17E-04	0,0022	0,0022
N70	TCAP	p.Arg106Cys	x	x	x	19	CH	Benigna	0,01958	0,0036	0,0084
	TNNI3	p.Arg162Trp	x	x	x	22,1	CH	Patogênica	-	8,18E-05	-
N72	SOS1	p. Ala708Thr	x	x	x	22,7	Noonan	Benigna	0,0053	2,30E-04	0,0042

¥ - Deleção de 3pb, levando a perda de um aminoácido glutamato. ‡ - Alteração em homozigose

Tabela 13: Alterações encontradas nas amostras reclassificadas como inconclusivas. Abreviações: Cardiomiopatia dilatada (CD); Cardiomiopatia hipertrófica (CH); Displasia arritmogênica de ventrículo esquerdo (DAVD); Morte súbita infantil (MSI); PROVEAN (Pr); SIFT (S); PolyPhen-2 (P2).

Amostra	Gene	Alteração	Deletéria			CADD	HGMD Assoc.	ClinVar	MAF		
			Pr	S	P2				ExAC	EVS	dbSNP
N4	<i>MYH6</i>	p.Ala1004Ser	x			22,4	CD	Patogênica	0,00098	9,23E-04	0,0006
N10	<i>SCN5A</i>	p.Gln692Lys				16,4	QT Longo	Duvidosa	0,00028	1,56E-04	0,0002
	<i>DSG2</i>	p.Val920Gly	x	x		8,2	DAVD	Duvidosa	0,0032	0,0038	0,0032
N11	<i>DSG2</i>	p.Ile293Val	x	x		12,9	DAVD	Benigna	0,06867	0,067	0,032
N12	<i>DSG2</i>	p.Ile293Val	x	x		12,9	DAVD	Benigna	0,06867	0,067	0,032
N13	<i>CSRP3</i>	p.Ala50Val	x	x		25,2	CD	-	2,47E-05	7,70E-05	-
N22	<i>PKP2</i>	p.Ser169Gly				2	DAVD	Duvidosa	0,0012	0,0016	0,0002
	<i>DSG2</i>	p.Val392Ile		x		23,5	DAVD	Patogênica	-	0,0015	0,001
	<i>MYH6</i>	p.Gly56Arg	x	x	x	12,7	CH?	Benigna	0,069	0,087	0,056
N24	<i>SCN5A</i>	p.Arg481Trp	x	x		22,4	-	Prov. Benigna	0,0011	0,003	0,0028
	<i>DTNA</i>	p.Pro586Leu	x	x	x	34	-	-	0,00013	3,84E-04	-
	<i>DES</i>	p.Val459Ile			x	21,2	CD	-	0,003	0,012	0,0098
N29	<i>NEBL</i>	p.Lys60Asn	x	x	x	24,3	CD	Benigna	0,038	0,005	0,014
N33	<i>DSP</i>	p.Glu1833Val	x	x	x	29,4	-	Benigna	0,0094	0,009	0,004
	<i>DES</i>	p.Val459Ile			x	21,2	CD	-	0,003	0,012	0,0098
N35	<i>DSC2</i>	p.Pro514Ala	x	x	x	23,7	-	-	-	-	-
N42	<i>MYBPC3</i>	p.Ala216Thr				13,7	CH?	Duvidosa	-	5,57E-04	0,001
N44	<i>PKP2</i>	p.Asp26Asn	x	x		21,6	DAVD	Prov. Benigna	0,014	0,004	0,003
N45	<i>MYBPC3</i>	p.Ala833Val	x		x	18,9	CH?	Benigna	0,0022	0,007	0,01
N48	<i>MYBPC3</i>	p.Ala833Val	x		x	18,9	CH?	Benigna	0,0022	0,007	0,01
N56	<i>DSP</i>	p.Glu1740Lys	x	x		10,8		Duvidosa	0,0012	9,23E-04	0,0004
	<i>DSG2</i>	p.Ile293Val	x	x		12,9	DAVD	Benigna	0,068	0,067	0,032
N58	<i>RYR2</i>	p.Arg298Cys	x	x	x	24,6	-	-	-	-	-
	<i>NEBL</i>	p.Tyr89*	-	-	-	35		Duvidosa	0,0012	0,0013	0,0006
	<i>DSG2</i>	p.Ile293Val		x	x	12,9	DAVD	Benigna	0,06867	0,067	0,032
N61	<i>RYR2</i>	p.Thr2504Met	x	x	x	33	DAVD	-	-	-	-
	<i>NEBL</i>	p.Lys60Asn	x	x	x	24,3	CD	Benigna	0,038	0,005	0,014
	<i>SCN5A</i>	p.Ser524Tyr	x	x	x	23,4	MSI	Duvidosa	0,005	0,011	0,014
	<i>TMPO</i>	p.Arg690Cys		x		24,5	CD	Duvidosa	0,015	6,15E-04	0,014
N62	<i>CAV3</i>	p.Thr78Met			x	27,2	QT Longo	Patogênica	0,003	0,0043	0,002
	<i>ANK2</i>	p.Glu3931Lys	x	x		24	Arritmia	Duvidosa	0,0026	0,003	0,0008
	<i>PKP2</i>	p.Thr482Met		x		12,11	DAVD	Prov. Benigna	0,0023	0,0037	0,0038
N63	<i>SCN5A</i>	p.Pro656Leu	x	x	x	16,5	-	Prov. Benigna	0,00039	0,0011	0,003
N64	<i>SCN5A</i>	p.Leu618Phe	x	x		24,6	QT Longo	Duvidosa	0,0005	0,0021	0,0016
N65	<i>MYH6</i>	p.Ala1004Ser	x			22,4	CD	Patogênica	0,00098	9,23E-04	0,0006
	<i>DSG2</i>	p.Glu713Lys			x	11,2	DAVD	Benigna	0,05	-	0,026
N71	<i>DES</i>	p.Val459Ile			x	21,2	CD	-	0,003	0,012	0,0098

Nas amostras positivas, foram detectadas 36 alterações distintas em 22 genes, sendo 10 sarcoméricos (*TNNI3*, *TPM1*, *TTN*, *MYL3*, *MYL2*, *TNNT2*, *MYBPC3*, *MYH6*, *MYH7* e *ACTC1*), quatro do disco-Z (*FLNC*, *ACTN2*, *TCAP* e *NEXN*), dois do citoesqueleto (*VCL* e *ANK2*) e um relacionado com a Linha-M (*TRIM63*). Ainda, foram encontradas alterações em cinco genes relacionados à fenocópias (*SOS1*, *RAF1*, *PRKAG2*, *TTR* e *MAP2K2*). O gene que apresentou maior frequência foi o da Filamina C gama (*FLNC*), com 6 ocorrências (13,9%) referentes a 6 mutações distintas (Figura 9), seguido pelo gene *TRIM63* com 4 ocorrências. Dos sarcoméricos, o mais frequente foi o gene *TNNI3*, com quatro alterações distintas, todas previamente associadas à CH. O gene *TPM1* apresentou 3 ocorrências referentes à mesma alteração (p.Arg21Leu), e o gene *MYL2* apresentou duas ocorrências referentes à alteração “p.Gly29Asp”, sendo que em ambos os casos a alteração foi detectada em homozigose.

Figura 9: Frequência de aparecimento de alterações nos 19 genes com mutação detectada nas amostras positivas. *genes sarcoméricos.

Oito das 35 amostras positivas (22%) apresentaram mais de uma alteração potencialmente patogênica e três amostras apresentaram alterações potencialmente patogênicas em homozigose (8,5%). Quatro amostras (N39, N40, N41 e N46) apresentaram alterações em genes para os quais já haviam sido sequenciadas através do sequenciamento Sanger (*MYH7*, *MYBPC3* e *TNNT2*), porém sem detecção de alteração patogênica. Não foram encontradas alterações em genes responsáveis pelo controle de cálcio. No que diz respeito a tipo de mutação, foram detectadas 32 do tipo *missense*, duas *nonsense* e apenas uma deleção de 3pb, levando à perda de um aminoácido glutamato na proteína Titina (N40).

Assim, dado que no Brasil 50% dos casos⁸ dos casos-índice de CH apresentam alterações potencialmente patogênicas nos três genes mais associados à CH (*MYH7*, *MYBPC3* e *TNNT2*) e, neste trabalho, 31 de 72 amostras negativas (43%) puderam ser reclassificadas como positivas em outros genes associados, isso confere um ganho de 21,5% em uma eventual casuística sem diagnóstico molecular prévio, aumentando para 71,5% a capacidade de detecção de alterações potencialmente patogênicas.

Em relação as amostras consideradas inconclusivas, o que mais chama a atenção é a alta ocorrência de alterações previamente relacionadas à Displasia arritmogênica do ventrículo direito (DAVD), sendo 10 ao todo. O gene da Desmogleína (*DSG2*) foi o mais frequente, com 7 mutações detectadas (Figura 10) em 7 amostras, sendo que a alteração “p.Ile293Val” foi observada em 4 delas. Mutações em outros 2 genes relacionados à DAVD também foram

detectadas, sendo 3 alterações no gene da placofilina-2 (*PKP2*) e um no gene do receptor de rianoldina (*RYR2*).

A segunda doença com mais alterações previamente associadas nos casos inconclusivos foi a Cardiomiopatia dilatada (CD), com 8 amostras apresentando alterações. O gene mais frequente foi o da Desmina (*DES*), sendo detectado em 3 amostras, porém todas acometidas pela mesma alteração (p.Val459Ile). A alteração “p.Ala1004Ser” do gene da cadeia pesada da α -miosina (*MYH6*) foi detectada em duas amostras, bem com a alteração “p.Lys60Asn” do gene da Nebulete (*NEBL*). Os genes da proteína LIM muscular (CSRP3) e da Timopoetina (TMPO) apresentaram uma alteração cada.

O gene da Subunidade alfa do canal de sódio tipo 5 (*SCN5A*) foi o segundo mais frequente, apresentando 5 ocorrências relativas a alterações distintas. Duas delas (p.Gln692Lys e p.Leu618Phe) estão descritas como associadas a Síndrome de QT longo e uma (p.Ser524Tyr) à morte súbita infantil. Quatro amostras apresentaram alterações duvidosas em relação à patogenicidade para CH, sendo três delas no gene *MYBPC3* e uma no gene *MYH6*. Por fim, 7 alterações sem descrição prévia foram detectadas, porém todas em genes não associados à CH (*SCN5A*, *DTNA*, *DSP*, *DSC2* e *RYR2*).

Figura 10: Frequência de aparecimento de alterações nos 16 genes com mutação detectada nas amostras inconclusivas.

5. DISCUSSÃO

5.1 Desempenho analítico do *pipeline*

Neste trabalho foram explorados três principais aspectos da aplicação do SNG e painéis genéticos no diagnóstico molecular: a performance do processo de enriquecimento de genes-alvo; o desempenho analítico da detecção de variantes, com o propósito de se identificar o balanço ideal entre uma boa sensibilidade, especificidade e uma baixa taxa de falsas descobertas; e, por fim, o ganho diagnóstico obtido com o uso de um painel genético ampliado para cardiomiopatias hereditárias. Para tanto, nós optamos pelo uso de um software comercial e avaliamos a capacidade do mesmo em analisar grandes quantidades de dados, sem a necessidade de interferências de programas externos nos pontos centrais da análise. Também desenhamos um processo de validação analítica utilizando amostras previamente sequenciadas por Sanger e um material referência do projeto HapMap (NA12878), seguindo a crescente tendência no uso desse tipo de material em validações^{41,53}.

De forma concordante com outras aplicações^{54,55}, a plataforma Ion Torrent PGM™ apresentou uma boa performance de sequenciamento, sendo capaz de gerar uma quantidade satisfatória de dados de boa qualidade (Phred $\geq$ Q20). As baixas taxas de policlonalidade contribuíram para que todas as corridas gerassem uma boa fração de resultados brutos. Dessa forma, mesmo com uma etapa de tratamento de sequências e filtragem de bases de baixa qualidade, a maioria das sequências geradas (93,83%) pôde ser utilizada no processo de mapeamento, e uma boa fração dessas foi mapeada em regiões-alvo (93,54%).

Esses valores estão diretamente relacionados com a boa performance do processo de enriquecimento. A cobertura média nas regiões-alvo se mostrou muito satisfatória (250x), uma vez que os valores de cobertura impactam diretamente na qualidade e confiabilidade da detecção de variantes⁵⁶. Todas as amostras processadas apresentaram uma boa cobertura, apesar da variabilidade inter-amstral observada (Figura 6). Entretanto, tal variabilidade parece não ter influenciado nos resultados de cobertura, dado que em média 95,2% das bases interrogadas tiveram cobertura mínima de 10x. Assim, isso deve estar mais relacionado com vieses de pipetagem e quantificação, como previamente observado⁴⁰, bem como à diferentes desempenhos entre as corridas.

Da mesma forma, a sensibilidade da chamada de variantes apresentou valores praticáveis tanto nas amostras providas do Painel CH-Sanger (96,5%) quanto para a NA12878 (94,7% na replicata NA12878_2), e a análise proposta dentro dos cinco cenários de cobertura e FAV conferiram robustez aos resultados finais. É perceptível a diferença nos resultados de sensibilidade ocasionada pela alteração de limiares de cobertura e FAV, o que levou a definição do ponto de 10x de cobertura/ 20%FAV como o mais sensível, ainda assim mantendo valores aceitáveis de especificidade. Os valores de reprodutibilidade aumentaram em paralelo aos de sensibilidade indicando que a identificação de alterações genéticas ocorreu de forma concordante entre todas as replicatas da NA12878. Valores de reprodutibilidade inter e intra-ensaio se mostraram próximos, indicando uma uniformidade no processamento dessas amostras. Notavelmente, nossos valores de reprodutibilidade com a

amostra NA12878 se mostraram muito próximos ao de valores reportados em validações de exoma completo⁵⁷. Tais resultados reiteram a importância de uma análise ampla de parâmetros analíticos para detecção de variantes, tais como aqui apresentados.

A correlação entre a queda da estringência de limiares e aumento de sensibilidade é, de certa forma, esperada, porém deve ser ajustada cautelosamente na tentativa de se evitar o aparecimento de falsas descobertas. Nosso desenho experimental com a amostra NA12878 foi um procedimento fundamental para traçar a ocorrência de FPs. Apesar do ponto mínimo de estringência apresentar as maiores taxas, tanto a validação inter como intra-ensaio apresentaram baixos valores de FPs. Das sete ocorrências mostradas nas Tabelas 10 e 11, quatro (2:179568916 T/G, 6:7580724 A/G, 6:7572262 A/G e 6:7584618 G/T) foram dentro de regiões de homopolímero, confirmando um viés da plataforma com o sequenciamento de tais regiões⁵⁸. Entretanto, nenhum FP foi encontrado dentro dos três genes mais associados à CH, tanto nas amostras do Painel CH-Sanger, como na validação completa do painel. Esse é um fato de grande importância dado que no Brasil esses genes dizem respeito à 50% dos casos com diagnóstico molecular positivo⁸.

É válido analisar NCs de forma diferente de FPs. Apesar de ambos não terem sido considerados como variantes legítimas, a análise separada dessas duas ocorrências permite o discernimento sobre diferentes problemas na chamada de variantes. O aparecimento de NCs indica um problema na identificação de genótipos, uma vez que a variante é verdadeira, porém foi detectada no genótipo errado. Em relação aos FPs, a ocorrência desses é

mais problemática uma vez que não são variantes legítimas e, dependendo da localização e do tipo de variante, podem vir a ser um fator de confusão em análises posteriores. Tanto NCs como FPs não foram vistos como problemas substanciais desse trabalho. Nesse mesmo sentido, o valor de TFD se apresentou diminuto enquanto o valor de VPP se apresentou acima de 0,959 durante todo o processo de validação.

Nossos resultados mostram que o *pipeline* validado apresenta uma tendência a favorecer a chamada de SNVs (sensibilidade máxima de 96,7%), enquanto que para InDels a resolução dos resultados e o baixo número de representantes desse tipo de variante na validação não nos permite estabelecer o mesmo (sensibilidade de 28,5%). Nota-se que de todas as variantes consideradas possivelmente patogênicas, tanto em casos positivos como inconclusivos, apenas uma é do tipo InDel. Somente InDels com 1pb de tamanho foram detectadas no processo de validação, porém a deleção encontrada na amostra N40 sugere um potencial do *pipeline* para detecção de deleções acima de 2pb. A princípio isso não impacta no uso do mesmo no diagnóstico molecular de CH, dado que a maioria das alterações patogênicas conhecidas são mutações pontuais ou InDels de 1pb⁵⁰.

Uma análise refinada das variantes não identificadas nos permitiu observar alguns pontos problemáticos no enriquecimento dos genes alvo. A principal razão para perda de variantes nas amostras do Painel CH-Sanger (Tabela 9) foi a ausência de cobertura bidirecional a qual ocorreu em 6 variantes, apesar da cobertura ser suficiente para identificação. Problemas de cobertura apareceram em duas variantes com menos de 10x devido a regiões

de alto conteúdo GC e uma não apresentava cobertura para alelo variante. Os mesmos problemas afetaram a identificação de variantes da amostra NA12878, sendo que a ausência de cobertura total também foi associada a regiões de alto conteúdo GC.

Analisando a distribuição de variantes não detectadas entre todas as corridas, nós não observamos uma correlação direta entre variantes perdidas e um baixo desempenho das corridas. Por exemplo, a corrida que apresentou o maior número de amostras com variantes perdidas foi a 2, mesmo com um desempenho de 1Gb (o segundo maior). A replicata de inter-ensaio da NA12878 que apresentou o maior número de variantes perdidas (12 de 14) foi a NA12878_3, incluída na corrida 11, a qual vem a ser a corrida com maior desempenho entre todas envolvidas na validação inter-ensaio. Esses dados sugerem que a falha na detecção de variantes está mais relacionada a regiões genômicas de difícil captura do que a variabilidade interamostral e desempenho das corridas. Tais áreas, uma vez identificadas, podem ser otimizadas através do redesenho das sondas responsáveis pelo enriquecimento. Outra forma utilizada na quantificação de regiões de baixa cobertura foi analisar a distribuição dos valores de cobertura média das regiões (Figura 5). Foram detectadas 28 regiões (1,6% do total) abaixo de 10x, sendo que dessas 8 (0,4%) já haviam sido indicadas pelo fabricante como parcial ou totalmente perdidas. Dessa forma, apenas 20 regiões se apresentaram problemáticas.

5.2 Ganho diagnóstico

O bom desempenho inicial do painel, e a definição de critérios de classificação de patogenicidade permitiu que obtivéssemos um valor considerável de ganho diagnóstico (21,5%), aumentando para 71,5% a probabilidade de positividade no teste molecular em casos-índices de CH, para a população brasileira. Estudos utilizando rastreamento genético para CH em probandos apresentam em torno de 30 a 38% de positividade, porém a maioria foca apenas ou majoritariamente em genes sarcoméricos^{14,59-61}.

Um achado interessante deste trabalho foi o fato dos dois genes mais frequentes não serem genes codificantes de miofibrilas sarcoméricas, mas sim referentes a componentes do disco-Z (*FLNC*) e da linha-M (*TRIM63*). O gene *FLNC* foi recentemente associado à CH em um estudo no qual foram identificadas 7 alterações genéticas segregando em 8 famílias com histórico da doença⁶², sendo que nas famílias portadoras dessas alterações havia uma alta incidência de morte súbita. Nenhuma das alterações encontradas nesse estudo foi detectada na nossa casuística, porém o fato de 5 das 6 alterações terem sido consideradas deletérias nos 3 algoritmos de predição, todas apresentarem valores de CADD acima de 15 e quatro delas não terem sido detectadas em nenhum dos bancos populacionais consultados, não descarta essas alterações como candidatas à causalidade. Além disso, com exceção da variante “p.Asp693Ala”, todas as outras alterações se encontravam dentro de domínios de interação proteica.

Já em relação às alterações encontradas em *TRIM63*, as duas alterações *missense* foram consideradas deletérias para os três algoritmos, se

apresentaram com baixa frequência em bancos de dados populacionais e obtiveram valores de CADD de 32. A alteração *nonsense* encontrada (p.Gln247*) apresenta associação com CH de acordo com o HGMD, porém a literatura é conflitante em relação à patogenicidade desta alteração. A primeira descrição de associação dessa alteração com CH foi feita por Chen et al. em 2012⁶³, em uma série de estudos funcionais mostrando perda de função da proteína, porém as famílias estudadas não eram grandes o suficiente para estudos de segregação. Em 2014, Ploski et al.⁶⁴ detectaram essa alteração em duas pessoas livres de CH. Há nos dados aqui apresentados amostras de dois indivíduos portadores dessa alteração, sendo um deles em homozigose. Assim, o segmento de familiares desses pacientes pode auxiliar no esclarecimento da patogenicidade e penetrância dessa alteração, dado que nenhum dos dois estudos prévios apresentou resultados de segregação familiar.

Uma observação interessante é o número de casos de dupla heterozigosidade encontrados. Enquanto no nosso estudo 22% das amostras positivas apresentaram mais de uma alteração genética potencialmente patogênica, outros estudos apontam para uma ocorrência de 6 a 8% na detecção desse tipo de caso^{59,60}. Já os casos em homozigose, apesar de raros, são descritos na literatura^{3,4} e aparecem em 3 casos aqui descritos (amostras N5, N34 e N55). A mesma diferença é observada em relação à detecção de casos de fenocópias. Nossos dados apresentam 5 casos (14,2%) nos quais foram detectadas alterações responsáveis por doenças de depósito, síndrome de Noonan ou doenças do espectro Noonan, muito provavelmente devido ao sub-diagnóstico desses casos, reiterando a importância de genes para

diagnóstico diferencial de CH em painéis como o aqui proposto. Uma vez identificados, pacientes portadores de doenças de depósito podem ser redirecionados para um tratamento específico. A fração de detecção desse tipo de caso representa 3% em outros estudos⁶⁰.

Um dos dados que mais chama à atenção diz respeito ao alto número de casos inconclusivos apresentando alterações genéticas descritas como causais para outras doenças cardíacas, principalmente relacionados com DAVD, ao invés de simplesmente duvidosas para CH. Esse fenômeno já foi observado em pacientes com CD⁶⁵, porém na literatura não é possível encontrar o mesmo para CH. Em um trabalho recente, Andreassen et al⁶⁶. questionaram a patogenicidade de alterações depositadas como causais para cardiomiopatias hereditárias (DAVD incluso) checando a frequência das mesmas no banco de dados ESP. Em um dos resultados, observaram que alterações nos genes DSP, DSG2 e PKP2 se encontravam acima da frequência populacional esperada, bem como em uma população controle, sugerindo revisão dessas alterações como possivelmente não patogênicas, ou como relacionadas a formas não monogênicas de DAVD.

Em concordância com esses achados, as alterações “p.Ile293Val” e “p.Glu713Lys” aqui encontradas no gene DSG2 e consideradas patogênicas pelo HGMD, apresentam frequência populacional acima de 1% em todos os bancos de dados. O mesmo podemos observar para as alterações “p.Lys60Asn” no gene *NEBL* e “p.Val459Ile” no gene *DES*, ambas consideradas patogênicas para CD no HGMD. Já a mutação “p.Gly56Arg” no gene *MYH6*, considerada duvidosa para CH, apresenta MAF acima de 5% em todas as

populações analisadas. Em uma rápida checagem no banco de dados de alterações específicas para DAVD (*ARVD/C Genetic Variants Database*⁶⁷) todas as alterações encontradas e indicadas como patogênicas no HGMD estavam classificadas como “patogenicidade incerta”, com exceção das variantes “p.Thr2504Met” e “p.Thr482Met”, nos genes RYR2 e PKP2, respectivamente, as quais não constavam no banco, indicando uma interpretação dúbia em relação ao impacto dessas alterações na patogênese da doença.

Dados como esses exemplificam a necessidade e a importância de uma estratégia de classificação de patogenicidade com múltiplas informações, uma vez que podem ser observadas incongruências entre informações depositadas na literatura, em bancos de dados de mutação e a frequência de aparecimento de determinada alteração em uma população controle. Com a crescente tendência de agregação de dados populacionais, a checagem da frequência de alterações potencialmente patogênicas tende a ser tornar uma prática de rotina em laboratórios com fim diagnóstico. No diagnóstico molecular de CH, a reanálise frequente de alterações genéticas com dados e ferramentas atualizadas já se mostrou eficaz na reclassificação de patogenicidade de mutações³, o que impacta diretamente no manejo de famílias e na confirmação de indivíduos em risco.

No contexto das nossas descobertas, tanto a reanálise frequente do impacto de alterações bem como o estudo de segregação familiar são dados fundamentais para a confirmação da potencial patogenicidade atribuída, em especial para os casos de dupla heterozigosidade e homozigose, os quais

apresentam um cenário diferenciado em relação à maioria dos casos de heterozigosidade simples. Por fim, consideramos também que a constante revisão dessas alterações é um fator de fundamental importância no ajuste dos valores de frequência de detecção de alterações patogênicas e no ganho diagnóstico, podendo direcionar decisões sobre a inclusão ou exclusão de genes a serem analisados na nossa população.

6. CONCLUSÕES

Considerando o bom desempenho da plataforma de sequenciamento Ion Torrent PGM™, a eficiência apresentada pelo método de enriquecimento de genes-alvo e a capacidade do *pipeline* desenhado na detecção de alterações genéticas, concluímos que nossa estratégia se apresenta praticável no sequenciamento de genes relacionados à cardiomiopatias hereditárias, tendo potencial para propiciar uma boa margem de casos com diagnóstico molecular positivo. Ainda, concluímos que a aplicação dos critérios de decisão utilizados nesse trabalho se mostraram capazes de indicar as alterações com maior potencial de patogenicidade, bem como apontar inconsistências entre as informações contidas em bancos de dados e a frequência de alterações observadas na população geral.

7. ANEXOS

Anexo A: Genes incluídos no painel e suas respectivas associações. Abreviaturas: Cardiomiopatia dilatada (CD); Cardiomiopatia hipertrófica (CH); Displasia arritmogênica ventricular direita (ARVD); Não compactação do ventrículo esquerdo (NCVE); Cardiomiopatia restritiva (CMR); Doenças do espectro Noonan (DEN); Síndrome de Noonan (SN); *CD com ocorrência de distrofia muscular de Emery–Dreifuss

Gene	Crom	Associação	Gene	Crom	Associação
<i>ACTN2</i>	1	CD	<i>CBL</i>	11	DEN
<i>LMNA</i>	1	CD	<i>CRYAB</i>	11	CD
<i>NEXN</i>	1	CD, CH	<i>CSRP3</i>	11	CD, CH
<i>NRAS</i>	1	SN	<i>HRAS</i>	11	DEN
<i>PSEN2</i>	1	CD	<i>MYBPC3</i>	11	CH, NCVE
<i>RYR2</i>	1	DAVD	<i>ABCC9</i>	12	CD
<i>TNNT2</i>	1	CD, CH	<i>KRAS</i>	12	SN
<i>TRIM63</i>	1	CH	<i>MYL2</i>	12	CH
<i>DES</i>	2	CD, DAVD	<i>PKP2</i>	12	DAVD
<i>SOS1</i>	2	SN	<i>PTPN11</i>	12	SN
<i>TTN</i>	2	CD, CH	<i>TMPO</i>	12	CD
<i>CAV3</i>	3	CH	<i>MYH6</i>	14	CD, CH
<i>MYL3</i>	3	CH	<i>MYH7</i>	14	CD, CH, NCVE
<i>RAF1</i>	3	SN	<i>PSEN1</i>	14	CD
<i>RPSA</i>	3	DAVD	<i>SYNE2</i>	14	CD*
<i>SCN5A</i>	3	CD	<i>TGFB3</i>	14	DAVD
<i>TMEM43</i>	3	DAVD	<i>ACTC1</i>	15	CD, CH, NCVE
<i>TNNC1</i>	3	CD, CH	<i>MAP2K1</i>	15	DEN
<i>ANK2</i>	4	CH	<i>SPRED1</i>	15	DEN
<i>MYOZ2</i>	4	CH	<i>TPM1</i>	15	CD, CH, NCVE
<i>SLC25A4</i>	4	CH	<i>GAA</i>	17	Doença de Pompe
<i>MYOT</i>	5	CD	<i>JUP</i>	17	CD, DAVD
<i>SGCD</i>	5	CD	<i>TCAP</i>	17	CD
<i>DSP</i>	6	DAVD	<i>DSC2</i>	18	DAVD
<i>EYA4</i>	6	CD	<i>DSG2</i>	18	DAVD
<i>PLN</i>	6	CD, CH	<i>DTNA</i>	18	NCVE
<i>SYNE1</i>	6	CD	<i>TTR</i>	18	Amiloidose Cardíaca
<i>BRAF</i>	7	DEN	<i>CALR3</i>	19	CH
<i>FLNC</i>	7	CD	<i>MAP2K2</i>	19	DEN
<i>PRKAG2</i>	7	CH	<i>TNNI3</i>	19	CD, CH, CMR
<i>FKTN</i>	9	CD	<i>JPH2</i>	20	CH
<i>BAG3</i>	10	CD	<i>MYLK2</i>	20	CH
<i>LDB3</i>	10	CD	<i>EMD</i>	X	CD*
<i>NEBL</i>	10	CD	<i>FHL1</i>	X	CH
<i>RBM20</i>	10	CD	<i>GLA</i>	X	Doença de Fabry
<i>SHOC2</i>	10	DEN	<i>LAMP2</i>	X	Doença de Danon
<i>VCL</i>	10	CD, CH	<i>TAZ</i>	X	CD, CH, NCVE

8. REFERÊNCIAS BIBLIOGRÁFICAS

1. Maron BJ, Maron MS. Hypertrophic cardiomyopathy. *Lancet*. 2013;381(9862):242-255.
2. Maron BJ, Maron MS, Semsarian C. Genetics of hypertrophic cardiomyopathy after 20 years: Clinical perspectives. *J Am Coll Cardiol*. 2012.
3. Das K J, Ingles J, Bagnall RD, Semsarian C. Determining pathogenicity of genetic variants in hypertrophic cardiomyopathy: importance of periodic reassessment. *Genet Med*. 2014;16(4):286-293.
4. Wang Y, Wang Z, Yang Q, Zou Y, Zhang H, Yan C, Feng X, Chen Y, Zhang Y, Wang J, Zhou X, Ahmad F, Hui R, Song L. Autosomal Recessive Transmission of MYBPC3 Mutation Results in Malignant Phenotype of Hypertrophic Cardiomyopathy. *PLoS One*. 2013;8(6).
5. Jarcho JA, McKenna W, Pare JA, Solomon SD, Holcombe RF, Dickie S, Levi T, Donis-Keller H, Seidman JG, Seidman CE. Mapping a gene for familial hypertrophic cardiomyopathy to chromosome 14q1. *N Engl J Med*. 1989;321:1372-1378.
6. Amberger J, Bocchini C a, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*. 2009;37(Database issue):D793-D796.
7. Frey N, Luedde M, Katus H a. Mechanisms of disease: hypertrophic cardiomyopathy. *Nat Rev Cardiol*. 2012;9(2):91-100.
8. Marsiglia JDC, Credidio FL, de Oliveira TGM, Reis RF, Antunes MDO, de Araujo AQ, Pedrosa RP, Barbosa-Ferreira JMB, Mady C, Krieger JE, Arteaga-Fernandez E, Pereira ADC. Screening of MYH7, MYBPC3, and TNNT2 genes in Brazilian patients with hypertrophic cardiomyopathy. *Am Heart J*. 2013;166(4):775-782.
9. Fatkin D, Seidman CE, Seidman JG. Genetics and disease of ventricular muscle. *Cold Spring Harb Perspect Med*. 2014;4(1):a021063.
10. Watkins H, Ashrafian H, Redwood C. Inherited cardiomyopathies. *N Engl J Med*. 2011;364(17):1643-1656.
11. Marian AJ. Hypertrophic cardiomyopathy: from genetics to treatment. *Eur J Clin Invest*. 2010;40(4):360-369.
12. Rodríguez JE, McCudden CR, Willis MS. Familial hypertrophic cardiomyopathy: basic concepts and future molecular diagnostics. *Clin Biochem*. 2009;42(9):755-765.
13. Lopes LR, Rahman MS, Elliott PM. A systematic review and meta-analysis of genotype-phenotype associations in patients with hypertrophic cardiomyopathy caused by sarcomeric protein mutations. *Heart*. 2013;99(24):1800-1811.
14. Van Driest SL, Ommen SR, Tajik AJ, Gersh BJ, Ackerman MJ. Yield of genetic testing in hypertrophic cardiomyopathy. *Mayo Clin Proc*. 2005;80(6):739-744.
15. Ingles J, Doolan a, Chiu C, Seidman J, Seidman C, Semsarian C. Compound and double mutations in patients with hypertrophic cardiomyopathy: implications for genetic testing and counselling. *J Med Genet*. 2005;42(10):e59.
16. Ho CY. Hypertrophic cardiomyopathy in 2012. *Circulation*. 2012;125(11):1432-1438.

17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74(12):5463-5467.
18. Disponível em: Human Genome Project Completion: Frequently Asked Questions. <http://www.genome.gov/11006943>. Acesso em: Janeiro, 2015
19. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature*. 2011.
20. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. 2014;15(2):256-278.
21. Check Hayden E. Technology: The \$1,000 genome. *Nature*. 2014;507(7492):294-295.
22. Shendure J, Lieberman Aiden E. The expanding scope of DNA sequencing. *Nat Biotechnol*. 2012;30(11):1084-1094.
23. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet*. 2011.
24. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet*. 2013;14(March 2013):295-300.
25. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet Med*. 2014;(August):1-8.
26. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, Bekheirnia R, Leduc MS, Kirby A, Pham P, Scull J, Wang M, Ding Y, Plon SE, Lupski JR, Beaudet AL, Gibbs RA, Eng CM. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *From Dep Mol Hum Genet N Engl J Med*. 2013;369:1502-1511.
27. Xuan J, Yu Y, Qing T, Guo L, Shi L. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Lett*. 2013;340(2):284-295.
28. Meder B, Haas J, Keller A, Heid C, Just S, Borries A, Boisguerin V, Scharfenberger-Schmeer M, Stähler P, Beier M, Weichenhan D, Strom TM, Pfeufer A, Korn B, Katus H a, Rottbauer W. Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies. *Circ Cardiovasc Genet*. 2011;4(2):110-122.
29. Mook ORF, Haagmans M a, Soucy J-F, van de Meerakker JB a, Baas F, Jakobs ME, Hofman N, Christiaans I, Lekanne Deprez RH, Mannens MM a M. Targeted sequence capture and GS-FLX Titanium sequencing of 23 hypertrophic and dilated cardiomyopathy genes: implementation into diagnostics. *J Med Genet*. 2013;50(9):614-626.
30. Glotov AS, Kazakov S V., Zhukova E a., Alexandrov A V., Glotov OS, Pakin VS, Danilova MM, Tarkovskaya I V., Niyazova SS, Chakova NN, Komissarova SM, Kurnikova E a., Sarana AM, Sherbak SG, Sergushichev A a., Shalyto A a., Baranov VS. Targeted next-generation sequencing (NGS) of nine candidate genes with custom AmpliSeq in patients and a cardiomyopathy risk group. *Clin Chim Acta*. 2015.
31. Gómez J, Reguero JR, Morís C, Martín M, Alvarez V, Alonso B, Iglesias S, Coto E. Mutation Analysis of the Main Hypertrophic Cardiomyopathy Genes Using Multiplex

- Amplification and Semiconductor Next-Generation Sequencing. *Circ J*. 2014;78(12):2963-2971.
32. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*. 2011;12(9):628-640.
 33. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs R a, Hurles ME, McVean G a. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073.
 34. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, Nickerson D a, Bamshad MJ, Akey JM. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013;493:216-220.
 35. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet Med*. 2008;10(4):294-300.
 36. Weiss MM, Van der Zwaag B, Jongbloed JDH, Vogel MJ, Brüggewirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp C a L, van Slegtenhorst M a, van den Wijngaard A, Waisfisz Q, Nelen MR, van der Stoep N. Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories. *Hum Mutat*. 2013;34(10):1313-1321.
 37. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15(9):733-747.
 38. Miller TE, You L, Myerburg RJ, Benke PJ, Bishopric NH. Whole blood RNA offers a rapid, comprehensive approach to genetic diagnosis of cardiovascular diseases. *Genet Med*. 2007;9(1):23-33.
 39. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol*. 2014;32(3):246-251.
 40. Li X, Buckton AJ, Wilkinson SL, John S, Walsh R, Novotny T, Valaskova I, Gupta M, Game L, Barton PJR, Cook S a, Ware JS. Towards clinical molecular diagnosis of inherited cardiac conditions: a comparison of bench-top genome DNA sequencers. *PLoS One*. 2013;8(7):e67744.
 41. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding K V, Zehnbaue B a, Agarwala R, Bennett SF, Chen B, Chin ELH, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmüller U, Gunesman SJ, Hegde MR, Johnson PLF, Kasarskis A, Kulkarni S, Lenk T, Liu CSJ, Manion M, Manolio T a, Mardis ER, Merker JD, Rajeevan MS, Reese MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol*. 2012;30(11):1033-1036.
 42. Feliubadaló L, Lopez-Doriga A, Castellsagué E, Del Valle J, Menéndez M, Tornero E, Montes E, Cuesta R, Gómez C, Campos O, Pineda M, González S, Moreno V, Brunet J, Blanco I, Serra E, Capellá G, Lázaro C. Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. *Eur J Hum Genet*. 2013;21(August):864-870.

43. Disponível em: USeq SourceForge. <http://sourceforge.net/projects/useq/>. Acesso em: Julho, 2014
44. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One*. 2012;7(10).
45. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. 2001:863-874.
46. Adzhubei I a, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7(4):248-249.
47. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.
48. Disponível em: Home - SNP - NCBI. <http://www.ncbi.nlm.nih.gov/snp/>. Acesso em: Janeiro, 2015
49. Disponível em: ExAC Browser. <http://exac.broadinstitute.org/>. Acesso em: Janeiro, 2015
50. Stenson PD, Ball E V, Mort M, Phillips AD, Shiel J a, Thomas NST, Abeysinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat*. 2003;21(6):577-581.
51. Disponível em: ClinVar. <http://www.ncbi.nlm.nih.gov/clinvar/>. Acesso em: Janeiro, 2015
52. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178-192.
53. Linderman MD, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, Mahajan M, Shah H, Kasarskis A, Schadt EE. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics*. 2014;7(1):20.
54. Chan M, Ji SM, Yeo ZX, Gan L, Yap E, Yap YS, Ng R, Tan PH, Ho GH, Ang P, Lee ASG. Development of a next-generation sequencing method for BRCA mutation screening: a comparison between a high-throughput and a benchtop platform. *J Mol Diagn*. 2012;14(6):602-612.
55. Costa JL, Sousa S, Justino A, Kay T, Fernandes S, Cirnes L, Schmitt F, Machado JC. Nonoptical massive parallel DNA sequencing of BRCA1 and BRCA2 genes in a diagnostic setting. *Hum Mutat*. 2013;34(4):629-635.
56. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15(9):733-747.
57. Linderman MD, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, Mahajan M, Shah H, Kasarskis A, Schadt EE. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Med Genomics*. 2014;7(1):20.
58. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput Biol*. 2013;9(4).

59. Andersen PS, Havndrup O, Hougs L, Sørensen KM, Jensen M, Larsen LA, Hedley P, Thomsen ARB, Moolman-Smook J, Christiansen M, Bundgaard H. Diagnostic yield, interpretation, and clinical utility of mutation screening of sarcomere encoding genes in Danish hypertrophic cardiomyopathy patients and relatives. *Hum Mutat.* 2009;30(3):363-370.
60. Alfares AA, Kelly MA, Mcdermott G, Funke BH, Lebo MS, Baxter SB, Shen J, McLaughlin HM, Clark EH, Babb LJ, Cox SW, Depalma SR, Ho CY, Seidman JG, Seidman CE, Rehm HL. Original Research Article Results of clinical genetic testing of 2 , 912 probands with hypertrophic cardiomyopathy : expanded panels offer limited additional sensitivity. *Genet Med.* 2015.
61. Erdmann J, Daehmlow S, Wischke S, Senyuva M, Werner U, Raible J, Tanis N, Dyachenko S, Hummel M, Hetzer R, Regitz-Zagrosek V. Mutation spectrum in a large cohort of unrelated consecutive patients with hypertrophic cardiomyopathy. *Clin Genet.* 2003;64(4):339-349.
62. Valdés-Mas R, Gutiérrez-Fernández A, Gómez J, Coto E, Astudillo A, Puente DA, Reguero JR, Lvarez V, Morís C, León D, Martín M, Puente XS, López-Otín C. Mutations in filamin C cause a new form of familial hypertrophic cardiomyopathy. 2014.
63. Chen SN, Czernuszewicz G, Tan Y, Lombardi R, Jin J, Willerson JT, Marian AJ. Human molecular genetic and functional studies identify TRIM63, encoding muscle RING finger protein 1, as a novel gene for human hypertrophic cardiomyopathy. *Circ Res.* 2012;111(7):907-919.
64. Ploski R, Pollak A, Müller S, Franaszczyk M, Michalak E, Kosinska J, Stawinski P, Spiewak M, Seggewiss H, Bilinska ZT. Does p.Q247X in TRIM63 cause human hypertrophic cardiomyopathy? *Circ Res.* 2014;114(2):10-13.
65. Elliott P, O'Mahony C, Syrris P, Evans A, Sorensen CR, Sheppard MN, Carr-White G, Pantazis A, McKenna WJ. Prevalence of desmosomal protein gene mutations in patients with dilated cardiomyopathy. *Circ Cardiovasc Genet.* 2010;3(4):314-322.
66. Andreassen C, Nielsen JB, Refsgaard L, Holst AG, Christensen AH, Andreassen L, Sajadieh A, Haunsø S, Svendsen JH, Olesen MS. New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur J Hum Genet.* 2013;(November 2012):918-928.
67. Lazzarini E, Jongbloed JDH, Pilichou K, Thiene G, Basso C, Bikker H, Charbon B, Swertz M, van Tintelen JP, van der Zwaag P a. The ARVD/C Genetic Variants Database: 2014 Update. *Hum Mutat.* 2015;36(4):403-410.