Universidade de Lisboa
Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Tratamento de dados de NGS para pesquisa de novas mutações associadas à Miocardiopatia Hipertrófica

Mónica Seabra Dourado Eusébio

Trabalho de Projecto

Mestrado em Bioestatística
2012

Universidade de Lisboa
Faculdade de Ciências

Departamento de Estatística e Investigação Operacional

Tratamento de dados de NGS para pesquisa de novas mutações associadas à Miocardiopatia Hipertrófica

Mónica Seabra Dourado Eusébio

Trabalho de projecto orientado por:
Prof. Doutora Lisete Maria Ribeiro de Sousa
Prof. Doutor Francisco Javier Enguita

Mestrado em Bioestatística
2012

Resumo

A Sequenciação de Nova Geração (NGS - *Next Generation Sequencing*) está a revolucionar a investigação na área da biomédica, contribuindo significativamente para o avanço da medicina personalizada. A NGS, apoiando-se nos conhecimentos da Estatística Bayesiana e da Bioinformática para a análise e tratamento dos dados que esta técnica origina, torna-se num excelente exemplo dos novos campos multidisciplinares emergentes da Ciência.

As novas potencialidades da NGS que a tornam tão atractiva para tantos novos projectos de investigação são a sua capacidade de misturar diversas amostras numa só leitura, com recurso a um sistema de código de barras, diminuindo custos e aumentando a rapidez da obtenção das amostras.

Com o intuito de descobrir potenciais novas mutações associadas à Miocardiopatia Hipertrófica, estudou-se uma coorte de indivíduos que possuiam o diagnóstico clínico da doença, mas que, no entanto, não apresentavam qualquer mutação exónica patogénica.

Tornam-se claras as vantagens desta nova metodologia, pois trazendo mais rapidez e permitindo uma análise a um maior número de genes, com custos reduzidos, torna possível a análise de todos os genes com associação conhecida e descrita na literatura. Neste projecto, obtiveram-se resultados positivos em relação a muitos genes já conhecidos que pela sequenciação padrão não tinham sido sequenciados ou detectados. A NGS demonstrou ter todo o potencial para tornar-se o novo método padrão de diagnóstico, sendo, por isso, necessário continuar aperfeiçoar e a melhorar a metodologias de tratamento e análise de dados provenientes desta nova técnica.

Foram descobertos sete genes (CAV3, GLA, LDB3, MYLK2, MYOZ2, PRKAG2 e VCL) que ainda não foram oficialmente associados à patologia, sendo que na literatura apenas são referidas mutações a nível exónico. Apenas o gene ANKRD1 detectado com alterações já foi descrito com mutações exónicas e intrónicas. O gene CAV3 surge como associado à doença oficialmente em 2010, num artigo de revisão.

Palavras-chave: Sequenciação de Nova Geração, Estatística Bayesiana, Bioinformática, Miocardiopatia Hipertrófica

Abstract

The Next-Generation Sequencing (NGS) is revolutionizing biomedical research, significantly contributing to the enrichment of health sciences' field, inclusively towards a personalized medicine. NGS takes advantage of Bayesian Statistics and Bioinformatics knowledge, in order to analyse and process data, which is originated by this technique. NGS is an excellent example of the new multidisciplinary fields of Sciences.

The innovative NGS's features, which make this technique so attractive for so many different research projects , are the possibility of mix several samples in just one read, using an associated barcode, which can reduce costs. Obtaining samples is quicker than with the standard method.

With the purpose of discover unknown mutations associated with Hypertrophic Cardiomyopathy, we studied a group of individuals, which, in spite of presenting the disease's symptoms, did not have any pathogenic exonic mutation with the standard method of sequentiation.

The advantages of this new methodology are quite clear. Accomplishing faster results and allowing analysis of a larger number of genes, with reduced costs, which allows to diagnose the patient concerning all genes described in the literature. In this project, we obtained positive results for many genes already known that were not sequenced or not detected by standard sequencing method. The NGS shown to have the necessary potential to become the new standard diagnostic method, and is therefore imperative to further refine and improve methods of treatment and analysis of data originated by this technique.

In our project, we found seven genes (CAV3, GLA, LDB3, MYLK2, MYOZ2, PRKAG2 and VCL) which were not officially associated with the disease yet and in the literature only exonic mutations were described. Only ANKRD1 gene was detected with both exonic and intronic mutations. CAV3 gene appears in 2010 as officially associated with the disease, in a review article.

Keywords: Next-generation Sequencing, Bayesian Statistics, Bioinformatics, Hypertrophic Cardiomyopathy

Agradecimentos

Seria impensável não dedicar uma parte da minha dissertação a um pequeno agradecimento a quem possibilitou que a mesma existisse.

Começo por referir a Prof. Lisete Sousa por ter-me encorajado a enviar uma candidatura a um anúncio do Instituto de Medicina Molecular (IMM) e à Prof. Carmo Fonseca por ter lido a candidatura.

Praticamente todo o meu trabalho foi realizado no IMM, sendo isto possível graças ao excelente acolhimento que tive por parte da Genomed. Não podia deixar de nomear algumas das pessoas que estiveram mais perto de mim, como a Rute Marcelino por ter estado sempre comigo, durante as minhas frustrações e vitórias iniciais; a Ana Rita Grosso pelo empurrão inicial; ao Pedro Eleutério e ao Daniel Guerreiro pela ajuda informática, sempre pedida de forma dramática, mas correspondida de forma paciente.

Ao Francisco Enguita, que aparecendo depois, conseguiu ser um apoio precioso, aconselhando-me e encorajando-me a continuar, mesmo quando parecia que me encontrava num beco sem saída.

Um agradecimento especial ao meu pai que sempre me pagou as propinas, mesmo sem saber muito bem aquilo que estudo ou faço.

A todos os que me acompanharam, o meu obrigada.

Mónica Eusébio

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Esta dissertação pretende incidir sobre um projecto de tratamento de dados de Sequenciação de Nova Geração, com o intuito de pesquisar novas mutações associadas à Miocardiopatia Hipertrófica numa coorte de indivíduos com diagnóstico clínico da doença, mas que por sequenciação padrão não revelaram qualquer mutação exónica patogénica. Uma das inovações deste projecto passa pelo estudo das zonas intrónicas dos genes, não só das exónicas. A análise irá incidir, não só nos genes já descritos na literatura, mas também nalguns possíveis candidatos. Aliado a esta vertente mais prática do projecto, haverá o aprofundamento das noções estatísticas que estão na base dos métodos bioinformáticos utilizados.

1.1 Sequenciação de Nova Geração

Após trinta anos de supremacia do método de Sanger, a Sequenciação de Nova Geração surge, por volta do ano de 2005, como uma técnica revolucionária para a investigação biomédica, sendo que a Bioinformática tem um papel de destaque na análise e interpretação dos dados obtidos. Os principais desafios desta nova prática prendem-se com a real identificação da mutação que provoca a doença dentro das numerosas variantes possíveis e com a falta de padrões na recolha, manipulação e transmissão da informação na investigação biomédica. Apesar destes obstáculos, a Sequenciação de Nova Geração consegue modificar a forma de fazer investigação ao permitir a realização de experiências, ao torná-las viáveis em termos técnicos e, talvez o mais importante, em termos económicos. [Lindblom e Robinson, 2011] [Voelkerding, Dames e Durtschi, 2009]

Na tecnologia de Sequenciação de Nova Geração utilizada neste projecto - *Illumina/Solexa Genome Analyzer*, mais especificamente o modelo Illumina HiSeq2000, os moldes de

ADN são lidos aleatoriamente ao longo do genoma com recurso à fragmentação de todo o genoma em fragmentos de pequena dimensão que serão ligados a adaptadores designados para leitura aleatória durante a síntese do ADN. Os resultados da sequenciação são apelidados de leituras curtas por terem cerca de cinquenta pares de bases. Esta técnica resulta em sinais luminosos que são descodificados de modo a determinar a base na sequência de ADN com o respectivo score de qualidade que nos descreve o quão provável é essa base ser a correcta. A precisão da atribuição de cada base é superior a 99.5%. A qualidade dos dados e o tamanho das leituras tornam esta técnica a mais utilizada em diversos projectos de sequenciação de genoma. As principais características que contribuem para o destaque desta nova metodologia são a possibilidade de ter diversas amostras misturadas numa só corrida (com a utilização de um código de barras, ou seja um conjunto de nucleótidos característico para cada indivíduo) e a rapidez da obtenção das leituras. [Nowrousian, 2010] [Zhang *et al.*, 2011]

Aprofundando um pouco o funcionamento desta metodologia em particular tem-se que o ADN ao ser fragmentado, vai ser posteriormente reparado nas pontas e que estas últimas serão ligadas a adaptadores. Após desnaturação, cada fragmento terá uma das terminações presa a um suporte sólido. Esse suporte tem a superfície totalmente coberta por adaptadores e adaptadores complementares. Cada cadeia isolada, presa por uma ponta, irá ligar-se ao adaptador complementar na superfície com a ponta anteriormente solta, criando uma estrutura em ponte por hibridação. Na mistura contendo os reagentes de amplificação de PCR, os adaptadores na superfície irão actuar como *primers* para a amplificação por PCR. Esta amplificação é necessária para intensificar o sinal luminoso, para termos uma detecção mais fiável das bases adicionadas. Vamos ter vários ciclos de PCR que irão formar grupos aleatórios de cerca de mil cópias de cada fragmento inicial de ADN na superfície sólida. A mistura de reacção adicionada para as reacções de sequenciação e síntese de ADN contém *primers*, quatro nuceótidos terminadores reversíveis, cada um identificado com uma tinta fluorescente distinta, e ADN polimerase. A cada incorporação de cada um dos nucleótidos terminadores no fragmento, a sua posição na superfície de suporte e a sua identificação pela tinta fluorescente é detectada por uma câmera própria para o efeito. A tinta é retirada dessa última base adicionada e o ciclo de síntese é repetido. No nosso projecto será repetido cerca de cinquenta vezes (as nossas leituras serão de um comprimento de cerca de cinquenta bases). Como teremos leituras *paired-end* a sequenciação será feita em ambas as extremidades do fragmento de ADN. Fazendo isto, podemos corrigir certos erros de leitura e aumentar a cobertura da amostra. [Ansorge, 2009] [Magi *et al.*, 2010] [Voelkerding *et al.*, 2010]

Para este projecto optou-se por sequenciamento alvo que consiste em sequenciar apenas

um conjunto de genes ou regiões, em vez de todo o genoma, poupando tempo e recursos. Esta metodologia é a escolhida quando se pretende descobrir ou validar variação genética na população. Uma das inovações deste trabalho foi o estudo das zonas intrónicas, para além das exónicas. Um intrão é a região que não codifica uma proteína no genoma e um exão é a zona que codifica. O intrão ainda que não codifique, é responsável pelo modo como a zona codificante é interpretada.

Um gene será toda a sequência de um ácido nucleico necessária à síntese de um produto génico, seja ele um polipéptido (conjunto de aminoácidos) ou um ARN. Outra noção importante é a de *splicing* alternativo, em que a partir do mesmo gene, isto é de um mesmo pré-ARN mensageiro transcripto, com diferentes combinações dos exões na formação do ARN mensageiro, codificar-se-á uma proteína distinta (tradução). Para além de intrões e exões, também encontramos zonas intergénicas de ADN não funcional. A estrutura de um gene é apresentada na Figura 1.1.[Hartwell *et al.*, 2008][Lodish et al., 2008]

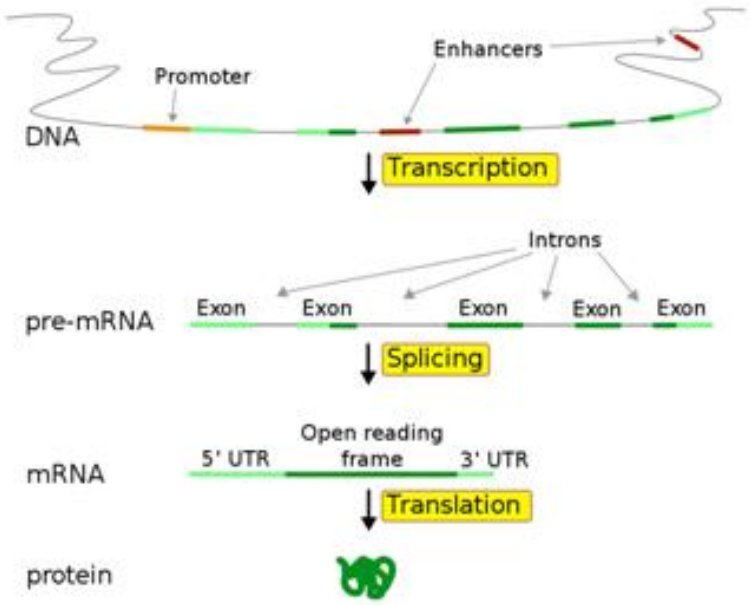

Figura 1.1: Estrutura de um gene. Numa primeira fase temos o ADN que por transcripção origina o pré-ARN. Este possui intrões e exões, permitindo que, por *splicing*, gere um determinado ARN mensageiro que por tradução vai codificar um proteína. O ARN é lido no sentido 5'UTR para 3'UTR. *Imagem retirada de http://en.wikipedia.org/wiki/File:Gene2-plain.svg*

1.2 Miocardiopatia Hipertrófica

A Miocardiopatia Hipertrófica é uma doença do miocárdio que é caracterizada por hipertrofia ventricular esquerda, com envolvimento predominante do septo interventricular. A sua prevalência está compreendida entre uma a cada mil pessoas a uma em cada quinhentas pessoas. Em 55% dos casos existe história familiar evidente, sendo a hereditariedade autossómica dominante[1]. A patologia tem uma grande variabilidade inter e intra-familiar, variando entre formas benignas e formas malignas (elevado risco de insuficiência cardíaca e morte súbita - que atinge um a dois porcento dos doentes). Os sintomas são muito variáveis e pouco específicos, variando entre: intolerância ao exercício físico, mesmo quando ligeiro; precordialgia (dor no peito); sensação de síncope (desmaio) eminente ou mesmo síncope; taquiarritmias (ritmo cardíaco aumentado) ou arritmias (ritmo cardíaco irregular); e sopro cardíaco. Na literatura estão descritos por volta de catorze genes associados à patologia. Na Tabela 1.1, na página 5, encontram-se enumerados os genes que serão analisados neste trabalho e respectivas localizações cromossómicas. [Seidman e Seidman, 2011] [Voelkerding *et al.*, 2010]

Figura 1.2: Coração de paciente com Miocardiopatia Hipertrófica (esquerda) e coração normal (direita). *Imagem retirada de:[Ho, 2011]*

Destes 26 genes em estudo, podem destacar-se os catorze genes que já se encontram descritos na literatura, bem como a proteína que codificam. Essa informação será resumida

[1] A designação autossómica refere-se à localização num cromossoma não sexual e dominante ao facto de bastar uma cópia do alelo afectado para a característica ser revelada.

Tabela 1.1: Genes em estudo relativos à Miocardiopatia Hipertrófica.

Nome do Gene	**Localização Cromossómica**
ACTC1	Cromossoma 15: 35,080,296 - 35,087,927 cadeia "reverse"
ACTN2	Cromossoma 1: 236,849,754 - 236,927,931 cadeia "forward"
ANKRD1	Cromossoma 10: 92,671,853 - 92,681,033 cadeia "reverse"
CAV3	Cromossoma 3: 8,775,486 - 8,883,492 cadeia "forward"
CSRP3	Cromossoma 11: 19,203,578 - 19,232,120 cadeia "reverse"
GLA	Cromossoma X: 100,652,791 - 100,662,913 cadeia "reverse"
LAMP2	Cromossoma X: 119,561,682 - 119,603,220 cadeia "reverse"
LDB3	Cromossoma 10: 88,426,549 - 88,495,825 cadeia "forward"
MYBPC3	Cromossoma 11: 47,352,957 - 47,374,253 cadeia "reverse"
MYH6	Cromossoma 14: 23,851,049 - 23,877,486 cadeia "reverse"
MYH7	Cromossoma 14: 23,881,947 - 23,904,927 cadeia "reverse"
MYL2	Cromossoma 12: 111,348,628 - 111,358,404 cadeia "reverse"
MYL3	Cromossoma 3: 46-899,362 - 46,923,659 cadeia "reverse"
MYLK2	Cromossoma 20: 30,497,111 - 30,422,492 cadeia "forward"
MYOZ2	Cromossoma 4: 120,056,939 - 120,108,944 cadeia "forward"
NEXN	Cromossoma 1: 78,354,198 - 78,409,580 cadeia "forward"
PLN	Cromossoma 6: 118,869,461 - 118,881,893 cadeia "forward"
PRKAG2	Cromossoma 7: 151,253,210 - 151,574,210 cadeia "reverse"
TCAP	Cromossoma 17: 37,820,440 - 37,822,808 cadeia "forward"
TNNC1	Cromossoma 3: 52,485,118 - 52,488,086 cadeia "reverse"
TNNI3	Cromossoma 19: 55,663,138 - 55,669,100 cadeia "reverse"
TNNT2	Cromossoma 1: 201,328,136 - 201,346,890 cadeia "reverse"
TPM1	Cromossoma 15: 63,334,831 - 63,364,111 cadeia "forward"
TTN	Cromossoma 2: 179,390,716 - 179,695,529 cadeia "reverse"
VCL	Cromossoma 10: 75,757,872 - 75,879,918 cadeia "forward"
JPH2	Cromossoma 20: 42,740,335 - 42,816,218 cadeia "reverse"

na Tabela 1.2, na página 6.

Tabela 1.2: Genes descritos na bibliografia relativos à Miocardiopatia Hipertrófica. *Informação retirada de: [Cirino e Ho, 2011].*

Nome do Gene	**Proteína**
ACTC1	Actina, músculo cardíaco
ACTN2	Alpha-Actina-2
CSRP3	Proteína 3 rica em cisteína e glicina, músculo
MYBPC3	Proteína C de ligação à Miosina, tipo cardíaco
MYH6	Cadeia pesada da Miosina isoforma alpha, músculo cardíaco
MYH7	Cadeia pesada da Miosina isoforma beta, músculo cardíaco
MYL2	Cadeia 2 leve reguladora da Miosina, isoforma ventricular/músculo cardíaco
MYL3	Polipéptido leve 3
TCAP	Teletonina
TNNC1	Troponina C, músculo lento e músculo cardíaco
TNNI3	Troponina I, músculo cardíaco
TNNT2	Troponina T, músculo cardíaco
TPM1	Tropomiosina 1, cadeia alpha
TTN	Titina

Conhecendo a estrutura do músculo cardíaco, torna-se mais intuitivo perceber a importância destes genes em específico. Na Figura 1.4 encontra-se uma imagem da estrutura muscular e na Figura 1.5 a representação da estrutura de um sarcómero.

Figura 1.3: visto ao microscópio electrónico de transmissão. Trata-se de um corte longitudinal. É visível a estrutura dos sarcómeros e um disco intercalado. *Imagem retirada de: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_publishing_group/documents/image/wtdv033034.jpg*

O músculo cardíaco apresenta um estriado cruzado. Uma das suas características é possuir discos intercalados, a atravessar transversalmente o tecido muscular, em intervalos irregulares (visível na Figura 1.3). Estes discos vão ter um papel relevante na contracção muscular. [Junqueira e Carneiro, 2003]

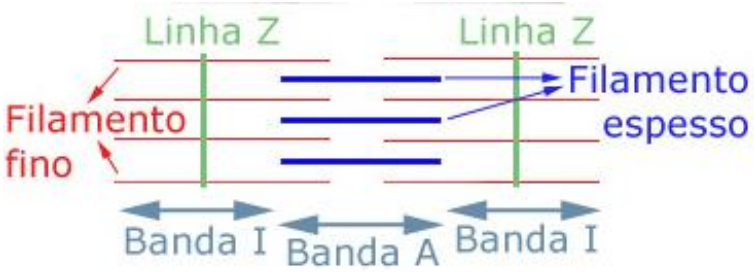

Figura 1.4: unidade responsável pela contracção muscular. *Imagem retirada de: http://pt.wikipedia.org/wiki/Ficheiro:Sarcómero.jpg*

Os discos intercalados possuem, nas porções transversais, uma membrana especializada que serve de âncora para os filamentos de actina dos terminais dos sarcómeros. No fundo representam as hemi-bandas Z do sarcómero. As bandas I são compostas por actina e as bandas A por miosina. A titina está representada pela linha vertical verde, na Figura 1.4. As fibras musculares possuem milhares de sarcómeros. [Junqueira e Carneiro, 2003]

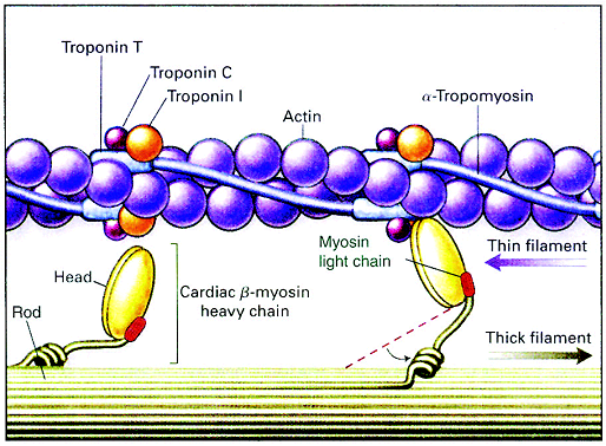

Figura 1.5: e respectiva organização. *Imagem retirada de: http://edoc.hu-berlin.de/dissertationen/kabaeva-zhyldyz-2002-11-11/HTML/kabaeva-ch1.html*

Na Figura 1.5, está representado de modo esquemático o processo de contracção do múculo. Temos então que quando o músculo se encontra em repouso, a cabeça de miosina se encontra ligada a uma molécula de ATP (Adenosina Tri-Fosfatada, energia da célula). Ao perder um fosfato inorgânico, hidrolisado pela cabeça da miosina (por acção do enzima ATPase), o ATP passa a ADP (Adenosina Di-Fosfatada) o que faz com que passe a existir afinidade entre a miosina e a actina que se ligarão formando um complexo. Dá-se a libertação do ADP e do fosfato inorgânico, bem como de energia. Isto deformará a cabeça da miosina, o que provocará um movimento que empurrará a actina para trás do filamento de miosina, assim a actina estará bem dentro da banda A. Deste modo, há contracção muscular. O sistema só reinicia quando o complexo se separa, após união entre a miosina e uma nova molécula de ATP, que irá repor a estrutura original da cabeça da miosina, preparando-a para um novo ciclo. Se não existir ATP disponível o ciclo não recomeça, tratando-se de rigidez muscular extrema, caso do *rigor mortis* que ocorre após a morte. [Junqueira e Carneiro, 2003]

É importante conhecer se um indivíduo possui ou não a doença, mesmo que esta não se manifeste em termos de sintomas e não haja propriamente uma cura. Sabendo que tipo de mutação a provoca pode adequar-se uma terapêutica mais eficaz a cada paciente, prevenindo complicações e diminuindo os sintomas. Os doentes com Miocardiopatia Hipertrófica podem ser sempre assintomáticos ou com sintomas ligeiros e passar a sintomas mais graves e complicações.

Capítulo 2

Aprofundamento dos Métodos Estatísticos

Neste capítulo será feito o aprofundamento da metodologia aplicada neste projecto, de forma a explicar as noções que estão na base dos programas e técnicas utilizadas. O capítulo será dividido em várias secções, seguindo a ordem pela qual o procedimento prático é realizado, podendo apresentar subsecções quando existem alternativas de metodologia. Tanto os diferentes tipos de ficheiros gerados como os programas utilizados serão abordados.

2.1 O formato *.fasta*

Este formato consiste numa sequência de nucleótidos. Permite atribuir um nome à sequência e, eventualmente, comentários. No Exemplo 1 será apresentado um pequeno excerto do genoma de referência utilizado no projecto (humano, *hg19*) que se encontra em *.fasta* referente ao cromossoma 1.

Exemplo 1.

```
>chr1
TGTTTTGACTATCCCCTCCACCCTCATCGCATTCGATATGGAGAGTAGGT
GGTATTAGGGAGATAACTTACTTAGAAAGGTACTTTCTCTGAATGGTGTA
TAGTTGACGATAGCCGATATGAGGGAAGAAAATACATAAGAGGACAAAAT
AGAATGCCCAGAAAGCTTTAGAAAATAATAGAAGACAGAAAGAAAAACAT
GATTATGGAAGAAGGATTAAGGTTGATGATGAGAGAAAGGGGACACTGAATTT
```

É neste formato que será sempre utilizado o genoma de referência, contra o qual será

feito o alinhamento.

2.2 O formato *.fastq*

O formato *.fastq* tem algumas variantes, por essa razão a explicação irá incidir sobre a variante correpondente à *Illumina/Solexa*. Sendo semelhante ao formato *.fasta* anteriormente descrito, utiliza a codificação PHRED para os scores de qualidade com uma janela de variação de 0 a 40, ainda que os scores PHRED variem entre 0 e 62. O *software* PHRED lê as sequências, determina a base e associa um score de qualidade a cada uma. A qualidade em escala PHRED (Q_{PHRED}) é estimada em termos de probabilidade de erro P_e, a partir da seguinte expressão $Q_{PHRED} = -10 \times \log_{10}(P_e)$. Como ilustração apresenta-se o Exemplo 2 referente a um dos indivíduos em estudo.[Cock *et al.*, 2010]

Exemplo 2.

```
@HWI-ST177_211 _C03PTACXX:6:1101:1330:1966♯CGTT\1
ATCACGCAATGCAGAAGAGAAAGNCAAGAAGGCCATCACTGATGTA
+
DADB:BA8<E≪C≪F:9FG?A*♯11?)C;F3D)0*00?*99/9)
@ HWI-ST177_211_C03PTACXX:6:1101:1405:1979♯CTTT\1
CAAATGGAATCGAATGGAATCACNGAACAGAATCGAATGGAACAAT
```

Observando o Exemplo 2, vemos o símbolo @ a iniciar a linha de título com a identificação da sequência e uma descrição opcional que pode conter o nome do instrumento, comprimento de leituras, etc. As bases de ADN podem ser apresentadas com os C, G, A, T habituais (representando respectivamente Citosina, Guanina, Adenina e Timina), com um N - quando não se sabe que base é, ou outros símbolos da lista da IUPAC (União Internacional da Química Pura e Aplicada) explicados na Tabela 2.1. [Deorowicz e Grabowski, 2011]

Tabela 2.1: Lista de Símbolos da IUPAC (retirado de *http://www.cisred.org/rat1.1/iupac_symbols _help*).

Símbolo	**Significado**
R	G ou A (purina)
Y	T ou C (pirimidina)
K	G ou T
M	A ou C

Símbolo	Significado
S	G ou C
W	A ou T
B	G, T ou C
D	G, A ou T
H	A, C ou T
V	G, C ou A

Todas as amostras do projecto vieram no formato *.fastq*, sendo que o nosso trabalho se inicia nesta fase.

2.3 Alinhamento ao genoma de referência

A primeira fase da metodologia consiste no alinhamento das amostra ao genoma de referência. Para este propósito, utilizou-se o programa *BWA*. Implementando a ferramenta de alinhamento de Burrows-Wheeler, este programa torna-se uma das melhores opções no que toca ao alinhamento de leituras curtas a uma sequência longa de referência, como é o caso das nossas amostras. Este programa também é vantajoso em relação a outros por permitir alinhamentos com leituras que apresentam lacunas (*gaps*) e por estimar a dimensão da biblioteca da sequenciação. [Langmead *et al.*, 2009] [Li e Durbin, 2009]

Os grandes problemas inerentes ao alinhamento prendem-se com a quantidade de informação que possuímos (o genoma humano consegue ter cerca de 2.2GB de tamanho) e com a eficiência/rapidez, do próprio método de alinhamento. [Langmead *et al.*, 2009]

A indexação de Burrows-Wheeler consiste numa permutação reversível de caracteres num texto. Imaginando que temos um determinado texto, T, vamos adicionar o símbolo \$ no seu extremo esquerdo. Considerando que \$ lexicograficamente vem em primeiro lugar que qualquer outro caracter presente em T, permuta-se $\$T$ tantas vezes quanto a sua dimensão adicionada uma unidade (correspondente ao \$ adicionado). Deste modo, vamos obter uma matriz em que cada linha terá uma das soluções das permutações (cada linha é uma permutação de $\$T$) e as linhas estarão organizadas por ordem alfabética. A coluna mais à direita será a transformação de Burrows-Wheeler que possui a mesma dimensão do texto original com \$. Este processo permite uma pesquisa que exige pouca capacidade de um computador, pela sua eficiência. [Langmead *et al.*, 2009]

Especificamente para o programa *BWA*, algumas alterações foram implementadas por Li e Durbin em 2009, para a transformação de Burrows-Wheeler poder ser utilizada em alinhamento de pequenas leituras de sequenciação contra um genoma de referência longo (como o caso do humano). Considerando Σ um alfabeto, sabemos que o símbolo \$ não consta do mesmo. Uma *string* $X = a_0a_1...a_{n-1}$ acaba sempre com o símbolo \$, ou seja

$a_{n-1} = \$$. Considere-se $X[i] = a_i$, com $i = 0, 1, ..., n-1$, como o i-ésimo caracter de X e $X_i = X[i, n-1]$ um sufixo de X. Vamos considerar um vector sufixo S, que irá corresponder a uma permutação dos valores de $0, ..., n-1$, então $S(i)$ conterá a posição inicial de cada X_i de partida. Neste caso, a transformação de Burrows-Wheeler, BWT (do inglês *Burrows-Wheeler Transformation*), é considerada aquela que possui $B[i] = \$$ quando $S(i) = 0$ e $B[i] = X[S(i)-1]$ caso contrário. A dimensão de X e de B é igual a n. A arrumação das permutações foi igualmente realizada por ordem alfabética, com a mesma consideração no que respeita a \$. Na Figura 2.1 encontra-se um fluxograma explicativo do processo da transformação de Burrows-Wheeler aplicada em *BWA* e na Figura 2.2, encontra-se um exemplo que ilustra o mesmo. [Li e Durbin, 2009]

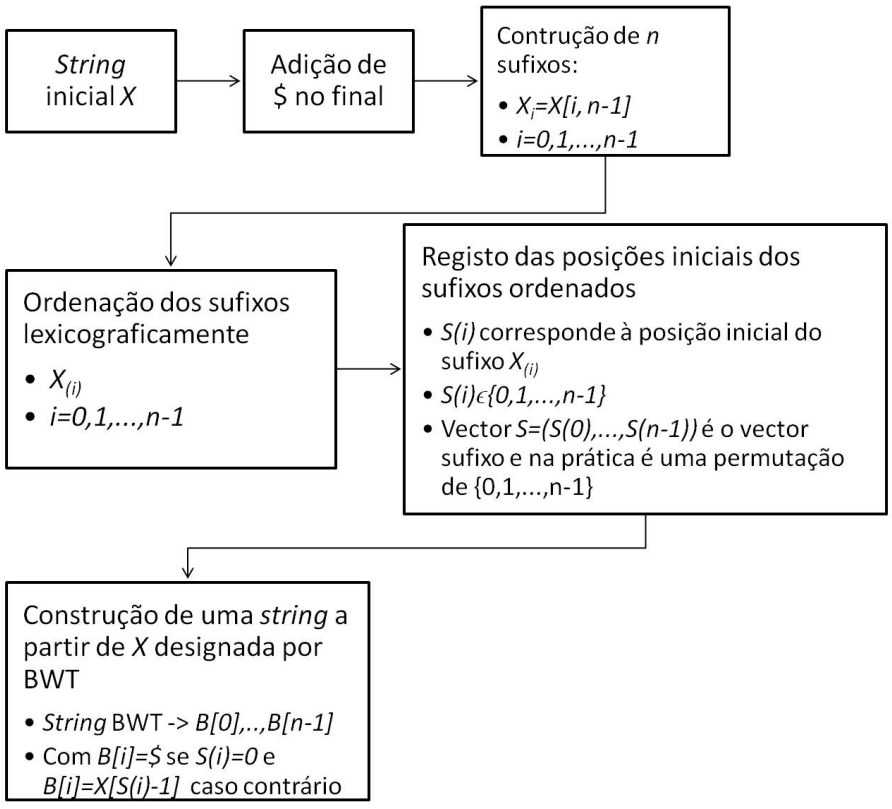

Figura 2.1: Fluxograma explicativo do funcionamento da transformação de Burrows-Wheeler aplicada ao programa *BWA*

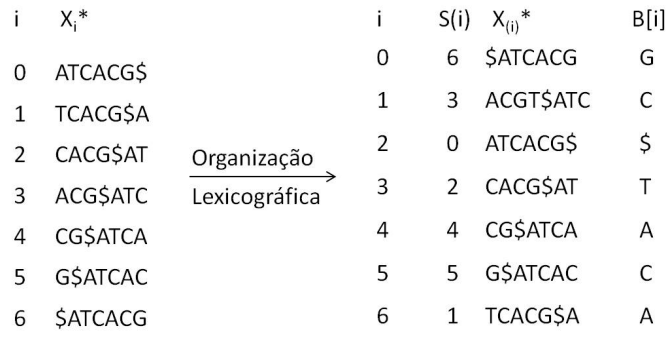

Figura 2.2: Explicação do funcionamento da transformação de Burrows-Wheeler aplicada ao programa *BWA*. Com base na *string* $X = ATCACG\$$ (os primeiros seis carateres do Exemplo 2.), constroiem-se sete sufixos (X_i) seguidos da sequência restante ($a_0...a_{i-1}$), gerando sete *strings* X_i^*. Estas sete *strings* são, posteriormente, organizadas por ordem alfabética, $X_{(i)}^*$. Após esta organização, as posições dos primeiros caracteres formam o vector sufixo $S(i)$, neste caso (6,3,0,2,4,5,1). A transformação de Burrows-Wheeler será a concatenação dos últimos caracteres das *strings* já organizadas, neste caso $GC\$TACA$.

Tendo em conta que estamos a trabalhar com sequenciamentos realizados em *Illumina*, o programa sempre que encontra um N (base desconhecida, conforme explicado na secção 2.2, na página 10), atribui uma das quatro bases aleatoriamente. Isto pode levar à ocorrência de falsos positivos, no entanto não é preocupante dada a raridade da situação. Para cada alinhamento, é calculado um score de qualidade de alinhamento de este ser incorrecto para cada posição. Considera-se que a verdadeira base pode ser sempre encontrada no genoma de referência, o que não é propriamente verdadeiro, levando a uma sobrestimação do score, ainda que com um desvio relativamente pequeno. [Li e Durbin, 2009]

No final do alinhamento, vamos ficar com um ficheiro *.sam*, para prosseguir com a análise.

2.4 O formato *.sam*

O formato *.sam* (SAM - *Sequence Alignment Map*), gerado após o processo de alinhamento, guarda a informação do alinhamento das leituras com o genoma de referência. O formato contém um cabeçalho, iniciado pelo símbolo @ e uma secção relativa ao alinhamento. Esta última possui onze campos obrigatórios e um número variável de campos opcionais. Quando algum dos campos obrigatórios não possui um valor, é preenchido por um zero ou um *, conforme o local. As colunas são separadas por TAB. [Li *et al.*, 2009]

O Exemplo 3 ilustra o formato *.sam* e a Tabela 2.2 possui a explicação dos onze campos obrigatórios do formato.

Exemplo 3.

```
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTA *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

Tabela 2.2: Explicação do formato *.sam*

(retirado de *[Li et al., 2009]*).

Nome da coluna	**Descrição**
QNAME	Nome do par de leitura
FLAG	Explicação em termos de bits da informação da leitura
RNAME	Nome da sequência de referência
POS	Posição no alinhamento da base mais à esquerda
MAPQ	Qualidade do alinhamento (na escala PHRED)
CIGAR	Descrição do alinhamento
MRNM	Referência da leitura par ("=" se for a mesma que em RNAME)
MPOS	Posição no alinhamento da base mais à esquerda da leitura par
ISIZE	Tamanho da inserção
SEQ	Sequência na mesma cadeia que a referência
QUAL	Qualidade (baseado em PHRED)

2.5 O formato *.bam*

O formato *.bam* (BAM - *Binary Alignment Map*) é a versão comprimida em formato binário do ficheiro *.sam*. Encontra-se na forma binária e o seu aspecto não é compreendido pelo humano, apenas pelo computador. No entanto, o programa *SAMtools*, possui um comando (comando *tview*) que permite a visualização do alinhamento no formato *.bam*. A Figura 2.3, ilustra a o alinhamento na região do gene ACTN2 do indivíduo representado pela notação 02187A. A primeira linha (contendo números) corresponde às posições, a

segunda linha à sequência do genoma de referência, as linhas seguintes vão corresponder cada uma a uma leitura, em que os pontos (ou vírgulas) indicam que nessa posição a leitura coincide com a referência. Quando as leituras não coincidem com a referência e aparecem letras, estas seguem a nomenclatura descrita na secção 2.2, na Tabela 2.1, na página 10. A cor alaranjada indica-nos que a qualidade dessa leitura não é muito boa.

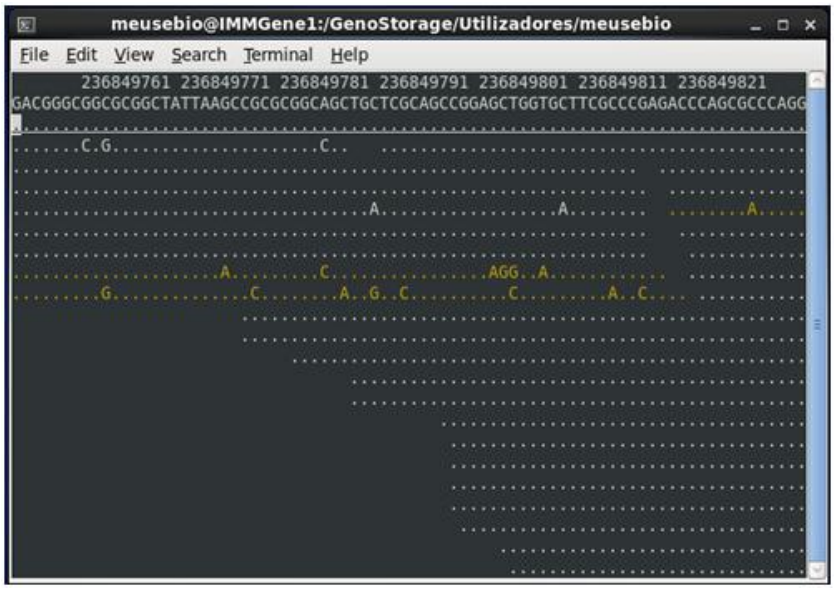

Figura 2.3: no programa *SAMtools* a partir do ficheiro no formato *.bam* do indivíduo 02187A, na região do gene ACTN2.

2.6 Determinação de SNP

Para a determinação dos SNP (alterações ao nível de um nucleótido), inserções (adição de um ou mais nucleótidos) ou delecções (perda de um ou mais nucleótidos), utilizou-se o programa *SAMtools* e testou-se o programa *GeMS*. [Li *et al.*, 2009] [You *et al.*, 2012]

2.6.1 Determinação recorrendo a *SAMtools*

Supõe-se que existe independência entre as leituras, mas este pressuposto pode ser desrespeitado quando ocorrem erros de alinhamento ou artefactos de PCR. Isto pode ser corrigido através de um esquema de pesos que tenha em consideração os erros de correlação ou recalibrando os scores de qualidade do alinhamento utilizando dados empíricos. [Nielsen *et al.*, 2011]

Supondo esta independência, a verosimilhança é o produto das verosimilhanças para as diferentes leituras. [Li, 2011]

Na Tabela 2.3 encontram-se descritas as notações que são mais usadas nesta secção de forma a facilitar a leitura da mesma.

Tabela 2.3: Notações mais habituais (adaptado de [Li, 2011]).

Símbolo	**Descrição**
n	Número de indivíduos
m_i	Número de cromossomas de cada indivíduo
M	Número total de cromossomas na amostra ($M = \sum_{i=1}^{n} m_i$)
D_i	Dados de sequenciação (nucleótidos e *scores* de qualidade) do i-ésimo indivíduo
g_i	Genótipo (número de alelos de referência) do i-ésimo indivíduo ($0 \leq g_i \leq m_i$)
ϕ_k	Probabilidade de observar k alelos de referência ($\sum_{k=0}^{M} \phi_k = 1$)
$\mathcal{L}_i(.)$	Função de verosimilhança de um parâmetro θ do i-ésimo indivíduo ($\mathcal{L}_i(\theta) = P\{D_i\|\theta\}$)
X	Contagem de alelos de referência num dado local para todos os indivíduos

Os alelos de referência são os alelos observados no indivíduo que são iguais aos do genoma de referência.

Considerando que temos uma amostra, os dados D são uma matriz contendo as bases com os respectivos scores de qualidade associados. Supondo que: num determinado local teremos k leituras, que as l primeiras bases ($l \leq k$) serão iguais às do genoma de referência e as restantes diferentes, e que a probabilidade de erro da j-ésima base da leitura é ϵ_j, para uma plóidia[1] de dois, partindo do pressuposto de independência entre erros, teremos: [Li, 2011]

$$\mathcal{L}(g) = \frac{1}{2^k} \prod_{j=1}^{l} [(2-g)\epsilon_j + g(1-\epsilon_j)] \prod_{j=l+1}^{k} [(2-g)(1-\epsilon_j) + g\epsilon_j] \tag{2.1}$$

em que g diz respeito ao número de alelos iguais aos do genoma de referência da amostra. [Li, 2011]

Definindo X como o número de alelos de referência nas amostras, a distribuição *a posteriori* de X é dada por: [Li, 2011]

[1] Plóidia refere-se ao conjunto de cromossomas por célula. Tratando-se de um projecto aplicado à espécie humana, a plóidia é dois; pois temos dois conjuntos de cromossomas homólogos (23 pares), ou seja somos diplóides (2n). [McCahill, 1996]

$$P\{X = k|D, \Phi\} = \frac{\phi_k P\{D|X = k\}}{\sum_l \phi_l P\{D|X = l\}} = \frac{\phi_k \mathcal{L}(k)}{\sum_l \phi_l \mathcal{L}(l)} \tag{2.2}$$

Note-se que D_i é uma matriz e $D = (D_1, ..., D_n)$ é uma matriz composta. $\mathcal{L}(k)$ diz respeito à verosimilhança da contagem alélica. A sua expressão é: [Li, 2011]

$$\mathcal{L}_D(k) = P\{D|X = k\} = \frac{1}{\binom{M}{k}} \sum_{g_1 \cdots g_n} \delta_{k,s_n(\mathbf{g})} \prod_{i=1}^{n} \binom{2}{g_i} \mathcal{L}_i(g_i) \tag{2.3}$$

onde $s_n(\mathbf{g})$ é o número total de alelos de referência na configuração genotípica $\mathbf{g}$ e $\delta_{k,l}$ a função delta de Kronecker que vale um quando $k = l$ e zero caso contrário. [Li, 2011]

Para a determinação das variantes, temos o forte conhecimento *a priori* que praticamente todos os nucleótidos das nossas leituras serão coincidentes com o genoma de referência. Recorrendondo à inferência bayesiana, podemos utilizar esta informação. Sendo ϕ_k a probabilidade de observar k alelos de referência entre os M cromossomas (haplótipos), com $k = 1, ..., M$ (M é o número de cromossomas presentes na amostra, como estamos a trabalhar com um indivíduo diplóide, cada amostra apresentará dois cromossomas (haplótipos), dois alelos, para cada característica, diferentes ou não). Define-se, por conveniência, $\Phi = (\phi_1, ..., \phi_k)$ que é a amostra AFS (espectro regional de frequência alélica) para M haplótipos. [Li, 2011]

A forma mais apropriada de estimar propriedades ao longo de múltiplos locais é utilizando o algoritmo EM-AFS. Considerando que temos L locais de interesse e que queremos calcular o AFS, então X_a, com $a = 1, ..., L$, é a variável aleatória que representa o número de alelos de referência no local a. Pode utilizar-se o algoritmo EM para achar o vector Φ que maximiza $P\{D|\Phi\}$, ou seja a probabilidade dos dados ao longo de todas as amostras com todos os locais condicionais a AFS. No entanto, a maior parte das vezes já existe um valor teórico para o AFS para cada local, proveniente de dados biológicos. A expressão de acordo com o algoritmo EM vem dada por: [Li, 2011]

$$\phi_k^{(t+1)} = \frac{1}{L} \sum_a P\{X_a = k|D, \Phi^{(t)}\} \tag{2.4}$$

A qualidade da variante é definida por $Q_{var} = -10 \log_{10} P\{X = M|D, \Phi\}$ e determina que o local é uma variante se Q_{var} for suficientemente grande. [Li, 2011]

Existem outras formas de determinar as variantes, mas foi esta a utilizada a partir do programa *SAMtools*.

2.6.2 Determinação recorrendo ao GeMS

Contrariamente à maioria dos programas para determinação de SNP, o *GeMS* toma em consideração os erros de preparação da amostra, para além dos erros de atribuição de

nucleótidos e de alinhamento. Este programa, recorre à maximização de verosimilhanças genotípicas e ao teste de Dixon para valores atípicos. [You *et al.*, 2012]

Para cada local genómico, s, iremos chamar D ao conjunto das informações do alelo alinhado. Note-se que D é diferente do apresentado na secção anterior, visto que não possui a informação relativa ao *score*. Sabendo que temos 10 genótipos possíveis dado que $G_1, G_2 \in \{A, C, T, G\}$, *GeMS* irá escolher $\text{argmax}_{G_1G_2} P(G_1G_2|D)$, para cada local s. Vamos considerar X_i o alelo observado na leitura i no local s e Y_i o verdadeiro alelo na leitura i no local s. [You *et al.*, 2012]

Considerando o genótipo G_1G_2, idealmente Y_i seria G_1 ou G_2. Como Y_i está sujeito a diversos erros (de preparação de amostra, entre outros), será uma variável aleatória latente que tem distribuição Categórica($\mathbf{p}^{G_1G_2}$), cujas probabilidades para cada caso são discriminadas na Tabela 2.4. O parâmetro p é a menor probabilidade de que Y_i iguala um alelo diferente do genótipo tido como referência. Uma distribuição Categórica é uma distribuição que atribui diferentes probabilidades a k possíveis valores de variável aleatória, com $k > 2$. [You *et al.*, 2012]

Tabela 2.4: Valores possíveis para $\mathbf{p}^{G_1G_2}$ para $Y_i \sim$Categórica($\mathbf{p}^{G_1G_2}$).
Informação retirada de: [You et al., 2012].

Modelo	G_1G_2	$p_A^{G_1G_2}$	$p_C^{G_1G_2}$	$p_G^{G_1G_2}$	$p_T^{G_1G_2}$
1	AA	$1-3p$	p	p	p
2	CC	p	$1-3p$	p	p
3	GG	p	p	$1-3p$	p
4	TT	p	p	p	$1-3p$
5	AC	$\frac{1-2p}{2}$	$\frac{1-2p}{2}$	p	p
6	AG	$\frac{1-2p}{2}$	p	$\frac{1-2p}{2}$	p
7	AT	$\frac{1-2p}{2}$	p	p	$\frac{1-2p}{2}$
8	CG	p	$\frac{1-2p}{2}$	$\frac{1-2p}{2}$	p
9	CT	p	$\frac{1-2p}{2}$	p	$\frac{1-2p}{2}$
10	GT	p	p	$\frac{1-2p}{2}$	$\frac{1-2p}{2}$

O programa irá, igualmente, utilizar a escala PHRED para a qualidade da determinação dos nucleótidos, B_i, bem como para a qualidade do alinhamento, M_i. Portanto, iremos ter $P(\text{determinação incorrecta da base}) = 10^{-0.1B_i}$ e $P(\text{alinhamento incorecto}) = 10^{-0.1M_i}$. Quando pretendemos calcular a precisão de um alelo alinhado esta é dada por, $w_i = min\{P(\text{determinação correcta da base}), P(\text{alinhamento correcto})\} = 1-10^{-0.1min\{B_i,M_i\}}$. [You *et al.*, 2012]

Dados o nucleótido e alinhamento correctos, podemos considerar $X_i = Y_i$ e propor a seguinte distribuição para X_i: $P(X_i = Y_i|Y_i) = w_i$, $P(X_i \neq Y_i|Y_i) = 1 - w_i$ e $P(X_i \neq Y_i, X_i = k|Y_i) = \frac{1-w_i}{3}$, com $k \in \{A, C, G, T\}$. [You *et al.*, 2012]

Considerando X_i independentes, com n_s sendo o número de leituras que abrange um local s ($i \in \{1, 2, ...n_s\}$), teremos:

$$\begin{aligned} L(\mathbf{p}^{G_1G_2}) = P(D|G_1G_2) &= \prod_{i=1}^{n_s} P(X_i = x_i) = \\ &= \prod_{i=1}^{n_s} \sum_{k\in\{A,C,G,T\}} [P(X_i = x_i|Y_i = k)]P(Y_i = k) = \\ &= \prod_{i=1}^{n_s} \sum_{k\in\{A,C,G,T\}} [w_i^{I(x_i=k)}(\frac{1-w_i}{3})^{I(x_i\neq k)}]p_k^{G_1G_2} \end{aligned}$$

Por definição, o programa utiliza uma distribuição *a priori* não informativa.

O genótipo consenso será baseado na probabilidade *a posteriori* e será denominado $\text{argmax}_{G_1G_2} P(G_1G_2|D) = \text{argmax}_{G_1G_2} L(\hat{\mathbf{p}}^{G_1G_2})$, com a seguinte ordem de estatísticas $P_{(1)} = min_{G_1G_2} P(G_1G_2|D) \leq P_{(2)} \leq ... \leq P_{(9)} \leq P_{(10)} = max_{G_1G_2} P(G_1G_2|D)$.

[You *et al.*, 2012]

Para um SNP ser determinado como tal, o programa utiliza o teste de Dixon para valores atípicos. Como cada local terá 10 probabilidades *a posteriori*, o valor da estatística de teste, Q, vem dado por: [Dixon, 1950] [You *et al.*, 2012]

$$Q = \frac{P_{(10)} - P_{(9)}}{P_{(10)} - P_{(2)}} \tag{2.5}$$

Este teste irá analizar a razão da diferença entre a maior probabilidade *a posteriori* e a segunda maior. O *output* do programa fornece o valor da maior probabilidade *a posteriori* para cada caso e o respectivo valor-p dos modelos genotípicos. Utiliza os valores da distribuição de Dixon tabelada para o cálculo do valor-p. Quando o valor-p é inferior a um nível de significância escolhido, a alteração é considerada um SNP. O valor-p é obtido com base no pacote `outliers` do programa **R** (http://cran.r-project.org/web/packages/outliers/index.html), criado por Lukasz Komsta, a partir do valor da estatística de teste. [You *et al.*, 2012]

A utilização deste programa não permite a progressão na análise dos dados, sendo necessário prosseguir com o *output* obtido no programa *SAMtools*.

2.7 O formato *.vcf*

Um ficheiro no formato .vcf guarda a informação relativa a polimorfismos no ADN (SNP, inserções e delecções) de forma compactada e de fácil acesso. O ficheiro é composto por uma zona de cabeçalho e outra de corpo. No Exemplo 5 (retirado do artigo de [Danecek *et al.*, 2011]) encontra-se uma representação de um ficheiro no formato *.vcf*, de forma a ilustrar o tipo de ficheiro obtido após a determinação das variantes.

No cabeçalho (todas as linhas começadas com ##) é especificado o formato, a data, o programa que gerou o ficheiro, o ficheiro de referência, todo esse género de identificações. No corpo do ficheiro, vamos ter uma primeira linha (iniciada com #) que identifica as colunas e as linhas seguintes com os dados. A primeira coluna é relativa ao número do cromossoma, a segunda à posição da primeira base da leitura, a terceira à identificação da variante, a quarta à sequência de referência, a quinta à alteração registada, a sexta ao score de qualidade, a sétima a informação relativamente ao filtro aplicado (se foi ou não considerada pelo filtro), as três últimas colunas só são utilizadas quando se analisam duas ou mais amostras em simultâneo. [Danecek *et al.*, 2011]

Exemplo 5.

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                FORMAT     SAMPLE1  SAMPLE2
1        1  .     ACG  A,AT    40  PASS    .                   GT:DP      1/1:13   2/2:29
1        2  .     C    T,CT    .   PASS    H2;AA=T             GT         0|1      2/2
1        5  rs12  A    G       67  PASS    .                   GT:DP      1|0:16   2/2:20
X      100  .     T    <DEL>   .   PASS    SVTYPE=DEL;END=299  GT:GQ:DP   1:12:.   0/0:20:36
```

2.8 Anotação de SNP

Esta parte final da metodologia tem como objectivo indicar, das variações detectadas anteriormente, quais as alterações que efectivamente provocam mudanças no fenótipo e qual a sua gravidade.

As amostras foram anotadas por dois programas diferentes (*snpEff* e *ANNOVAR*) de forma a encontrar a melhor alternativa em termos de interpretação de *output* final. Ambos os programas aceitam ficheiros no formato *.vcf* e anotam SNP, inserções e delecções.

2.8.1 Anotação utilizando *snpEff*

O programa *snpEff* utiliza as base de dados dos SNP do projecto 1000 Genomes (base de dados *hg19* do grupo NCBI e base de dados *GRCh37* do grupo Ensembl). Há pouca informação relativamente ao funcionamento do programa, dado que o artigo ainda não foi publicado. Em termos de descrição do efeito da variação é melhor que o *ANNOVAR*. O programa indica claramente o tipo de alteração que ocorreu, de que tipo é e categoriza o impacto. Gera um ficheiro que pode ser lido no Microsoft Excel^{TM}. [Cingolani, 2012]

2.8.2 Anotação utilizando *ANNOVAR*

A anotação deste programa é baseada em bases de dados pré-compiladas de anotação (do NCBI, UCSC Genome Browser e Ensembl). No final, apresenta um ficheiro que pode ser lido em Microsoft Excel^{TM}com toda a informação resumida. Em termos de nomenclatura do local de ocorrência da variação é melhor que o *snpEff*. [Wang, Li e Hakonarson, 2010]

As vantagens deste programa em relação a outros de anotação são o facto de servir para diversas plataformas de sequenciação, não necessitar de montar alinhamentos com as bases de dados e conseguir lidar com mutações sinónimas (redundantes, em que o resultado final é o mesmo) e a possibilidade de anotação de SNP desconhecidos (não descritos). A anotação pode ser relativa aos genes (a utilizada neste projecto) ou a zonas funcionais do genoma. O programa consegue filtrar e anotar variações que não estão descritas em bases de dados públicas. [Wang, Li e Hakonarson, 2010]

O programa exige que se converta o ficheiro em *.vcf* para um formato específico para uso interno no ANNOVAR (preparação do *input*) e a descarga de diversas bases de dados, mesmo que não utilize todas.

Capítulo 3

Análise dos Dados da Miocardiopatia Hipertrófica

Com o objectivo de descobrir novas mutações associadas à Miocardiopatia Hipertrófica, sequenciaram-se por Sequenciação de Nova Geração em *Harvard- Partners Center for Genetics and Genomics*, na Illumina HiSeq2000, os genes de uma coorte de dez pacientes com o diagnóstico clínico da doença, mas que não apresentavam qualquer mutação exónica patogénica por sequenciação padrão. Normalmente são testados apenas os cinco genes mais frequentes associados à doença, tanto pelo tempo que a sequenciação demora (podem ser vários meses) bem como pelos custos associados.

Os dados vieram em formato *.fastq*, já com o respectivo código de barras retirado e com as amostras separadas por indivíduos. O formato *.fastq* consiste na informação da sequência de nucleótidos da amostra com o respectivo score de qualidade para cada base. A explicação do formato é aprofundada no Capítulo 2, na secção 2.2, na página 10. [Cock *et al.*, 2010]

O código de barras trata-se de uma das inovações da Sequenciação de Nova Geração que permite numa só corrida ter diversas amostras de diferentes indivíduos, recorrendo à adição de uma sequência de oligonucleótidos característica de cada indivíduo. Esta sequência de identificação é retirada já na fase de análise dos dados, após a separação das amostras. Como os dados foram-nos fornecidos já sem código de barras, não foi necessário o passo referente ao tratamento do mesmo.

O servidor utilizado para a análise das amostra foi o IMMGene1, com as seguintes característica: 24 processadores Intel Xeon X5690, com 3.47GHz de velocidade cada e 1600.00MHz de bus, 94.5Gb de RAM com o sistema operativo Kernel Linux 2.6.32-71.29.1.el6.x86_64 com distribuição CentOS 6.2, pertencente ao Instituto de Medicina Molecular, da Faculdade de Medicina da Universidade de Lisboa.

A cada um dos indivíduos analisados foi associado um código identificativo. A inter-

pretação final dos resultados foi individual com base na amostra. A anotação das variações foi feita recorrendo a dois programas diferentes (*snpEff* e *ANNOVAR*) e a duas bases de dados distintas (*hg19* da UCSC (Universidade da Califórnia, Santa Cruz) em ambos os programas e *GRCh37.65* do EBI-EMBL em parceria com o Wellcome Trust Sanger Institute, do projecto Ensembl, apenas no último). A análise foi efectuada por indivíduo, ainda que a análise conjunta aumente a potência da determinação das variantes, pois em termos de diagnóstico o tratamento dos dados será sempre por indivíduo. [Flicek *et al.*, 2011] [Fujita *et al.*, 2011] [Li, 2011]

Apenas a anotação das variações é apresentada, pois esta é a única parte do tratamento dos dados que é analisada. No entanto toda a metodologia explicada no Capítulo 2 é aplicada, seguindo o fluxograma explicativo da Figura 3.1.

Figura 3.1: Fluxograma explicativo que ilustra a sequência aplicada no tratamento dos dados.

Para o programa *snpEff*, são apenas representados os dados relativos aos genes não descritos na literatura e com cobertura superior a 30 (uma variação é considerada relevante se possuir uma cobertura superior a 30). Quanto aos alinhamentos, temos o mesmo critério de apresentação. A partir das variações obtidas neste programa é que vamos comparar com as obtidas no programa *GeMS*, para sabermos se se tratam de possíveis candidatas a

SNP. Ainda que aparente ser uma inversão na ordem descrita, tanto na Figura 3.1 como no Capítulo 2, não o é, dado que o programa *GeMS* apenas apresenta a posição do gene de forma meramente numérica, portanto precisamos de obter os dados já anotados para poder confrontar com esse *output*.

Para a visualização dos alinhamentos nos locais onde ocorrem as variações em cada indivíduo foi utilizado o programa *IGV*[1], já com os ficheiros em formato *.bam* (explicação do formato no capítulo 2, secção 2.5, na página 14), utilizando *hg19* como genoma de referência. [Robinson *et al.*, 2011][Thorvaldsdóttir, Robinson e Mesirov, 2012]

Não é possível visualizar todo o gene em simultâneo no alinhamento, portanto optou-se por apresentar as zonas mais relevantes de cada gene e indivíduo. Opta-se por apresentar o alinhamento na fase final da interpretação dos dados, pois só nesta fase sabemos que zonas do alinhamento são mais importantes de visualizar e os alinhamentos foram agrupados por gene para sabermos até que ponto as alterações se encontram nos mesmos locais para todos os indivíduos com variações nesses genes, em vez de apresentar um alinhamento por gene e por indivíduo.

As variações não são correctamente anotadas por uma questão de confidencialidade, sendo apenas descrito o tipo de variação, que gene afecta, bem como o local funcional, e que tipo de efeito provoca. Quando encontramos o mesmo gene repetido, com a mesma cobertura e qualidade, mas com efeitos diferentes, trata-se da mesma variação com todos os efeitos previstos.

3.1 Amostra 02186A

Na Tabela 3.1, observa-se que nenhuma das duas mutações assinaladas pelo programa *ANNOVAR* é relevante, dado que a sua cobertura é extremamente baixa (apenas dois, quando deveria ser superior a 30). A que possui impacto trata-se de uma variação sinónima. Apenas um dos genes não se encontra na literatura (NEXN).

Tabela 3.1: Mutações detectadas no indivíduo 02186A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
MYBPC3	G por A	Exão	NA	NA	Sinónima
NEXN	adição de T	Intrão	101	NA	NA

Na Tabela 3.2 encontram-se os valores das verosimilhança do genótipo das alterações detectadas utilizando o programa *ANNOVAR* com a base de dados *hg19*, na escala PH-

[1] O programa *IGV - Integrative Genomics Viewer* foi utilizado apenas para a visualização dos dados já alinhados.

RED.

Tabela 3.2: Verosimilhança do genótipo em escala PHRED das alterações detectadas no indivíduo 02186A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Referência**	**Alteração e Referência**	**Alteração**
NEXN	Nenhum - 139	T - 0	TT - 86

Na Tabela 3.2 temos a representação dos genótipos possíveis para cada alteração em cada gene. Utilizando os valores dados na tabela, que se encontram numa escala PHRED, conseguimos as probabilidades de se obterem aqueles dados condicional a cada um dos três genótipos.

1. NEXN
   (a) $P(Dados|--) = 10^{-13.90} = 1.259 \times 10^{-14}$
   (b) $P(Dados|-T) = 10^{0} = 1$
   (c) $P(Dados|TT) = 10^{-8.60} = 2.512 \times 10^{-9}$

Neste amostra não temos valores do programa *GeMS*, pois, tratando-se esta alteração de uma inserção, não poderia ser um SNP.

Apenas o programa *ANNOVAR* fornece as informações relativamente à verosimilhança. Sabe-se pelo fórum **SEQanswers** que no gene NEXN é o genótipo TT que é o mais provável.

Na Tabela 3.3 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *hg19*.

Tabela 3.3: Mutações detectadas no indivíduo 02186A, utilizando o programa *snpEff* com a base de dados *hg19*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
ANKRD1	T por A	Intrão	61	53	Modificador
CAV3	T por G	Intrão	103	38	Modificador
CAV3	T por G	Transcripto	103	38	Modificador
GLA	G por A	Intrão	144	34	Modificador
LDB3	T por C	Intrão	133	70	Modificador
LDB3	T por C	5'UTR	133	70	Modificador
LDB3	T por C	*upstream*	133	70	Modificador

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
LDB3	A por G	Intrão	140	34	Modificador
LDB3	A por T	Intrão	157	40	Modificador
MYOZ2	T por C	Intrão	112	89	Modificador
PRKAG2	G por A	Intrão	139	100	Modificador
PRKAG2	A por G	Intrão	216	41	Modificador
VCL	A por G	Intrão	65	32	Modificador

Os 7 genes identificados na Tabela 3.3 não se encontram descritos na literatura.

Na Tabela 3.4 encontram-se os valores-p e respectivas conclusões retiradas do programa *GeMS*, relativamente aos dados fornecido pelo *snpEff*, com a base de dados *hg19* que serão os mesmos que para a base de dados *GRCh37.65*. Só são apresentados os valores-p, visto que o programa não fornece os valores da estística de teste.

Tabela 3.4: Valores-p das alterações detectadas no indivíduo 02186A - *snpEff* - *hg19* no programa *GeMS*.

Gene	Alteração	Valor-p	Conclusão
CAV3	T por G	0.0000	Possível SNP
GLA	G por A	0.0000	Possível SNP
LDB3	T por C	0.0000	Possível SNP
LDB3	A por G	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
MYOZ2	T por C	0.0000	Possível SNP
PRKAG2	G por A	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	A por G	0.0266	Possível SNP

Na Tabela 3.5 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.5: Mutações detectadas no indivíduo 02186A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
ANKRD1	T por A	Intrão	61	53	Modificador
CAV3	C por A	Intrão	103	38	Modificador

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
CAV3	C por A	Transcripto	103	38	Modificador
GLA	G por A	Intrão	144	34	Modificador
GLA	G por A	Transcripto	144	34	Modificador
GLA	G por A	*upstream*	144	34	Modificador
LDB3	T por C	Intrão	133	70	Modificador
LDB3	T por C	*upstream*	133	70	Modificador
LDB3	A por G	Intrão	140	34	Modificador
LDB3	A por G	Transcripto	140	34	Modificador
LDB3	A por T	Intrão	157	40	Modificador
LDB3	A por T	Transcripto	157	40	Modificador
MYOZ2	T por C	Intrão	112	89	Modificador
PRKAG2	G por A	Intrão	139	100	Modificador
PRKAG2	G por A	Transcripto	139	100	Modificador
VCL	A por G	Intrão	65	32	Modificador
VCL	A por G	Transcripto	65	32	Modificador
VCL	T por C	Intrão	189	86	Modificador
VCL	T por C	*downstream*	189	86	Modificador
VCL	delecção de TG	Intrão	136	97	Modificador
VCL	delecção de TG	*downstream*	136	97	Modificador

Utilizando uma base de dados diferente na Tabela 3.5, surgem novamente 7 genes que não se encontram na literatura, com possíveis SNP.

3.2 Amostra 02187A

Utilizando o programa *ANNOVAR*, com a base de dados *hg19*, não foi detectada qualquer mutação nos genes em estudo neste projecto.

Na Tabela 3.6 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *hg19*.

Tabela 3.6: Mutações detectadas no indivíduo 02187A, utilizando o programa *snpEff* com a base de dados *hg19*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
ANKRD1	C por A	Intrão	120	38	Modificador
LDB3	G por C	Intrão	219	72	Modificador
LDB3	T por C	Intrão	134	83	Modificador

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
LDB3	A por T	Intrão	147	62	Modificador
LDB3	G por A	Intrão	152	62	Modificador
MYLK2	T por C	Factor de iniciação ganho	136	71	Baixo
MYOZ2	G por A	Intrão	150	78	Modificador
MYOZ2	T por C	Intrão	158	93	Modificador
PRKAG2	A por T	Intrão	152	66	Modificador
PRKAG2	C por T	Intrão	151	95	Modificador
PRKAG2	T por C	Intrão	113	48	Modificador
PRKAG2	T por C	Intrão	103	40	Modificador
PRKAG2	A por C	Intrão	119	31	Modificador
PRKAG2	A por G	Intrão	165	69	Modificador
VCL	C por T	Intrão	141	81	Modificador

Os 6 genes identificados na Tabela 3.6, não se encontram descritos na literatura.

Na Tabela 3.7 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *hg19*, com recurso ao programa *GeMS*.

Tabela 3.7: Valores-p das alterações detectadas no indivíduo 02187A - *snpEff* - *hg19* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
ANKRD1	C por A	0.0000	Possível SNP
LDB3	G por C	0.0000	Possível SNP
LDB3	T por C	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
LDB3	G por A	0.0000	Possível SNP
MYLK2	T por C	0.0000	Possível SNP
MYOZ2	G por A	0.0000	Possível SNP
MYOZ2	T por C	0.0000	Possível SNP
PRKAG2	A por T	0.0000	Possível SNP
PRKAG2	C por T	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	A por C	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	C por T	0.0000	Possível SNP

Na Tabela 3.8 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.8: Mutações detectadas no indivíduo 02187A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
ANKRD1	C por A	Intrão	120	30	Modificador
LDB3	G por C	Intrão	219	72	Modificador
LDB3	G por C	Transcripto	219	72	Modificador
LDB3	T por C	Intrão	134	83	Modificador
LDB3	T por C	Transcripto	134	83	Modificador
LDB3	A por T	Intrão	147	62	Modificador
LDB3	A por T	Transcripto	147	62	Modificador
MYLK2	T por C	Factor de iniciação ganho	136	71	Baixo

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
MYLK2	T por C	5'UTR	136	71	Modificador
MYOZ2	G por A	Intrão	150	78	Modificador
MYOZ2	T por C	Intrão	158	93	Modificador
PRKAG2	A por T	Intrão	152	66	Modificador
PRKAG2	A por T	Transcripto	152	66	Modificador
PRKAG2	A por T	*downstream*	152	66	Modificador
PRKAG2	A por T	*upstream*	152	66	Modificador
PRKAG2	C por T	Intrão	151	95	Modificador
PRKAG2	C por T	Transcripto	151	95	Modificador
PRKAG2	T por C	Intrão	113	48	Modificador
PRKAG2	T por C	Transcripto	113	48	Modificador
PRKAG2	T por C	*downstream*	113	48	Modificador
PRKAG2	T por C	Intrão	103	40	Modificador
PRKAG2	T por C	Transcripto	103	40	Modificador
PRKAG2	A por C	Intrão	119	31	Modificador
PRKAG2	A por C	Trancripto	119	31	Modificador
PRKAG2	A por C	*downstream*	119	31	Modificador
PRKAG2	A por G	Intrão	165	69	Modificador
PRKAG2	A por G	Transcripto	165	69	Modificador
VCL	C por T	Intrão	141	81	Modificador
VCL	C por T	Transcripto	141	81	Modificador
VCL	C por T	*upstream*	141	81	Modificador

As alterações detectadas com a base de dados *GRCh37.65*, apresentadas na Tabela 3.8, são as mesmas que com *hg19*, ainda que esta indique mais efeitos. Os 6 genes detectados possuem possíveis SNP.

3.3 Amostra 02188A

O programa *ANNOVAR* não detectou qualquer alteração nos genes em estudo.

Na Tabela 3.9 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *hg19*.

Tabela 3.9: Mutações detectadas no indivíduo 02188A, utilizando o programa *snpEff* com a base de dados *hg19*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
CAV3	C por A	Intrão	107	60	Modificador
CAV3	C por G	Intrão	141	38	Modificador
GLA	G por A	Intrão	158	30	Modificador
LDB3	G por A	*upstream*	142	90	Modificador
LDB3	A por G	Intrão	138	42	Modificador
LDB3	C por T	Intrão	139	41	Modificador
LDB3	A por G	Intrão	134	43	Modificador
LDB3	A por T	Intrão	123	54	Modificador
MYLK2	T por C	Factor de iniciação ganho	147	76	Baixo
MYLK2	T por C	5'UTR	147	76	Modificador
MYLK2	G por C	Intrão	69	131	Modificador
PRKAG2	T por C	Intrão	91	50	Modificador
PRKAG2	T por C	Intrão	93	47	Modificador
PRKAG2	T por C	Transcripto	93	47	Modificador
PRKAG2	T por C	*downstream*	93	47	Modificador
PRKAG2	A por C	Intrão	177	37	Modificador
PRKAG2	A por C	Transcripto	177	37	Modificador
PRKAG2	A por G	Intrão	218	38	Modificador
VCL	T por C	Intrão	57	76	Modificador

O programa *snpEff*, com a base de dados *hg19* (Tabela 3.9), apresenta variações significativas e modificadoras de efeito em 6 genes não descritos na literatura.

Na Tabela 3.10 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *hg19*, com recurso ao programa *GeMS*.

Tabela 3.10: Valores-p das alterações detectadas no indivíduo 02188A - *snpEff* - *hg19* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
CAV3	C por A	0.0000	Possível SNP
CAV3	C por G	0.0000	Possível SNP
GLA	G por A	0.0000	Possível SNP
LDB3	G por A	0.0000	Possível SNP
LDB3	A por G	0.0000	Possível SNP
LDB3	C por T	0.0000	Possível SNP

Gene	Alteração	Valor-p	Conclusão
LDB3	A por G	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
MYLK2	T por C	0.0000	Possível SNP
MYLK2	G por C	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	A por C	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP

Na Tabela 3.11 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.11: Mutações detectadas no indivíduo 02188A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
CAV3	C por A	Intrão	107	60	Modificador
CAV3	C por A	Transcripto	107	60	Modificador
CAV3	C por G	Intrão	141	38	Modificador
CAV3	C por G	Transcripto	141	38	Modificador
GLA	G por A	Intrão	158	30	Modificador
GLA	G por A	*upstream*	158	30	Modificador
GLA	G por A	*upstream*	158	30	Modificador
LDB3	G por A	*upstream*	142	90	Modificador
LDB3	A por G	Intrão	138	42	Modificador
LDB3	A por G	Transcripto	138	42	Modificador
LDB3	C por T	Intrão	139	41	Modificador
LDB3	C por T	Transcripto	139	41	Modificador
LDB3	A por G	Intrão	134	43	Modificador
LDB3	A por G	Trancripto	134	43	Modificador
LDB3	A por T	Intrão	123	54	Modificador
LDB3	A por T	Transcripto	123	54	Modificador
MYLK2	T por C	Factor de iniciação ganho	147	76	Baixo

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
MYLK2	T por C	5'UTR	147	76	Modificador
MYLK2	G por C	Intrão	69	131	Modificador
MYLK2	G por C	*upstream*	69	131	Modificador
PRKAG2	T por C	Intrão	91	50	Modificador
PRKAG2	T por C	Transcripto	91	50	Modificador
PRKAG2	T por C	Intrão	93	47	Modificador
PRKAG2	A por C	Intrão	177	37	Modificador
PRKAG2	A por G	Intrão	218	38	Modificador
PRKAG2	A por G	Transcripto	218	38	Modificador
VCL	T por C	Intrão	57	76	Modificador
VCL	T por C	*downstream*	57	76	Modificador

Com o programa *snpEff*, com ambas as bases de dados obtemos alterações significativas em 6 genes não descritos na literatura, todos com possíveis SNP.

3.4 Amostra 02189A

Na Tabela 3.12 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*.

Tabela 3.12: Mutações detectadas no indivíduo 02189A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
NEXN	A por C	Intrão	106	5	NA
NEXN	A por G	Intrão	93	9	NA
NEXN	inserção	Intrão	125	NA	NA

Na Tabela 3.12, observa-se que no gene NEXN, as duas primeiras substituições referem-se a alterações comuns, previstas em tabelas já publicadas. Quanto à última alteração, não contém informação quanto à cobertura.

Na Tabela 3.13 encontram-se os valores das verosimilhanças dos genótipos das alterações detectadas utilizando o programa *ANNOVAR* com a base de dados *hg19*, na escala PHRED.

Tabela 3.13: Verosimilhança do genótipo em escala PHRED das alterações detectadas no indivíduo 02189A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Referência**	**Alteração e Referência**	**Alteração**
NEXN	AA - 139	AC - 15	CC - 0
NEXN	AA - 123	AG - 0	GG - 31
NEXN	Nenhum - 163	-T - 0	TT - 103

1. NEXN

   (a) $P(Dados|AA) = 10^{-13.90} = 1.259 \times 10^{-14}$
   (b) $P(Dados|AC) = 10^{-1.50} = 3.162 \times 10^{-2}$
   (c) $P(Dados|CC) = 10^{0} = 1$

2. NEXN

   (a) $P(Dados|AA) = 10^{-12.30} = 5.012 \times 10^{-13}$
   (b) $P(Dados|AG) = 10^{0} = 1$
   (c) $P(Dados|GG) = 10^{-3.10} = 7.943 \times 10^{-4}$

3. NEXN

   (a) $P(Dados|--) = 10^{-16.30} = 5.012 \times 10^{-17}$
   (b) $P(Dados|-T) = 10^{0} = 1$
   (c) $P(Dados|TT) = 10^{-10.30} = 5.012 \times 10^{-11}$

Para o gene NEXN, no primeiro casos é a substituição, no segundo o genótipo AG e no último TT.

Recorrendo ao programa *GeMS*, é possível obter o valor-p para o valor da estatística de teste de Dixon para valores atípicos, utilizada neste contexto para verificar se as alterações se tratam efectivamente de SNP. Para a primeira alteração obtemos um valor-p de praticamente zero (apresentado no programa como 0.0000), o que nos indica que há evidência para rejeitar a hipótese nula de que não se trata de um SNP para qualquer um dos níveis de significância usuais. Quanto à segunda alteração, temos um valor-p de 0.0317, existindo evidência de que se trata efectivamente de um SNP a um nível de significância de 5%.

Na Tabela 3.14 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.14: Mutações detectadas no indivíduo 02189A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
ANKRD1	C por A	Intrão	150	40	Modificador
LDB3	A por T	Intrão	154	41	Modificador
LDB3	A por T	Transcripto	154	41	Modificador
LDB3	G por A	Intrão	136	39	Modificador
LDB3	G por A	Transcripto	136	39	Modificador
MYLK2	G por A	Intrão	146	98	Modificador
MYOZ2	T por C	Intrão	134	96	Modificador
PRKAG2	T por C	Intrão	65	52	Modificador
PRKAG2	T por C	Transcripto	65	52	Modificador
PRKAG2	T por C	*downstream*	65	52	Modificador
PRKAG2	T por C	Intrão	64	43	Modificador
PRKAG2	T por C	Transcripto	64	43	Modificador
PRKAG2	T por C	*downstream*	64	43	Modificador
PRKAG2	A por G	Intrão	185	49	Modificador
PRKAG2	A por G	Transcripto	185	49	Modificador
VCL	T por C	Intrão	222	89	Modificador
VCL	T por C	*downstream*	222	89	Modificador
VCL	delecção de TG	Intrão	31.5	85	Modificador
VCL	delecção de TG	*downstream*	31.5	85	Modificador

Como as base de dados *GRCh37.65* e *hg19*, fornecem os mesmos resultados, passarão a ser só apresentados os dados relativos à primeira base de dados, visto que nos indica mais informação relativamente aos efeitos previstos. Com o programa *snpEff* detectaram-se 6 genes não descritos na literatura com alterações significativas.

Na Tabela 3.15 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.15: Valores-p das alterações detectadas no indivíduo 02189A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
ANKRD1	C por A	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP

Gene	Alteração	Valor-p	Conclusão
LDB3	G por A	0.0000	Possível SNP
MYLK2	G por A	0.0000	Possível SNP
MYOZ2	T por C	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP

Todas as susbtituições detectadas por *snpEff*, são possíveis SNP.

3.5 Amostra 02190A

Na Tabela 3.16 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*. O gene MYBPC3 apresenta duas substituições com cobertura 91, no programa *ANNOVAR*, mas para além de serem sinónimas, uma delas é comum e está prevista em tabelas já publicadas.

Tabela 3.16: Mutações detectadas no indivíduo 02190A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
MYBPC3	duas substituições	Exão	225	91	Sinónimas

Na Tabela 3.17 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.17: Mutações detectadas no indivíduo 02190A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
LDB3	T por C	Intrão	155	92	Modificador
LDB3	T por C	*upstream*	155	92	Modificador
LDB3	T por C	*downstream*	155	92	Modificador
LDB3	C por T	Intrão	124	40	Modificador
LDB3	C por T	Transcripto	124	40	Modificador
LDB3	A por G	Intrão	134	38	Modificador

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
LDB3	A por G	Transcripto	134	38	Modificador
LDB3	A por T	Intrão	182	92	Modificador
LDB3	A por T	Transcripto	182	92	Modificador
PRKAG2	T por C	Intrão	172	32	Modificador
PRKAG2	T por C	Transcripto	172	32	Modificador
PRKAG2	A por G	Intrão	142	49	Modificador
PRKAG2	A por G	Transcripto	142	49	Modificador
VCL	T por C	Intrão	123	82	Modificador
VCL	T por C	*downstream*	123	82	Modificador

Para a amostra 02190A, na Tabela 3.17, são detectados 3 genes não descritos na literatura.

Na Tabela 3.18 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.18: Valores-p das alterações detectadas no indivíduo 02190A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
LDB3	T por C	0.0000	Possível SNP
LDB3	C por T	0.0000	Possível SNP
LDB3	A por G	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP

Todas as alterações detectadas nos genes são possíveis candidatos a SNP, de acordo com o programa *GeMS*.

3.6 Amostra 02191A

Na Tabela 3.19 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*. O programa *ANNOVAR* apenas detecta uma variação com cobertura não significativa.

Tabela 3.19: Mutações detectadas no indivíduo 02191A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
MYBPC3	T por C	Exão	19.8	2	Não Sinónima

Na Tabela 3.20 encontram-se os valores das verosimilhanças dos genótipos das alterações detectadas utilizando o programa *ANNOVAR* com a base de dados *hg19*, na escala PHRED.

Tabela 3.20: Verosimilhança do genótipo em escala PHRED das alterações detectadas no indivíduo 02191A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Referência**	**Alteração e Referência**	**Alteração**
MYBPC3	TT - 156	TC - 0	CC - 255

1. MYBPC3

   (a) $P(Dados|TT) = 10^{-15.60} = 2.512 \times 10^{-16}$
   (b) $P(Dados|TC) = 10^{0} = 1$
   (c) $P(Dados|CC) = 10^{-25.50} = 2.262 \times 10^{-26}$

Para o gene MYBPC3, o genótipo mais provável é o TC.

Na Tabela 3.21 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.21: Mutações detectadas no indivíduo 02191A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
LDB3	T por C	Intrão	154	84	Modificador
LDB3	T por C	*upstream*	154	84	Modificador
LDB3	A por G	Intrão	139	44	Modificador
LDB3	A por G	Transcripto	139	44	Modificador
LDB3	A por T	Intrão	152	40	Modificador
LDB3	A por G	Transcripto	152	50	Modificador
MYLK2	T por G	Intrão	12.3	69	Modificador
MYLK2	T por G	*upstream*	12.3	69	Modificador
MYOZ2	T por C	Intrão	118	92	Modificador
PRKAG2	G por A	Intrão	91	92	Modificador
PRKAG2	G por A	Transcripto	91	92	Modificador
PRKAG2	G por A	*downstream*	91	92	Modificador
PRKAG2	A por G	Intrão	140	31	Modificador
PRKAG2	A por G	Transcripto	140	31	Modificador
VCL	T por C	Intrão	70	54	Modificador
VCL	T por C	*downstream*	70	54	Modificador

O programa *snpEff* detectou 5 genes não descritos na literatura com alterações significativas, conforme apresentado na Tabela 3.21.

Na Tabela 3.22 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.22: Valores-p das alterações detectadas no indivíduo 02191A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
LDB3	T por C	0.0000	Possível SNP
LDB3	A por G	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
MYLK2	T por G	0.0000	Possível SNP
MYOZ2	T por C	0.0000	Possível SNP
PRKAG2	G por A	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP

Gene	Alteração	Valor-p	Conclusão
VCL	T por C	0.0000	Possível SNP

Todas as alterações foram consideradas possíveis candidatas a SNP pelo programa *GeMS*.

3.7 Amostra 02192A

Na Tabela 3.23 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*. O gene NEXN apresenta uma substituição com uma cobertura de 24 que puderá ser relevante.

Tabela 3.23: Mutações detectadas no indivíduo 02192A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
MYBPC3	G por A	Exão	NA	NA	Sinónima
NEXN	G por T	Intrão	13.2	24	NA
NEXN	A por G	Intrão	3.98	3	NA

Na Tabela 3.24 encontram-se os valores das verosimilhanças dos genótipos das alterações detectadas utilizando o programa *ANNOVAR* com a base de dados *hg19*, na escala PHRED.

Tabela 3.24: Verosimilhança do genótipo em escala PHRED das alterações detectadas no indivíduo 02192A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	Referência	Alteração e Referência	Alteração
NEXN	GG - 43	GT - 0	TT - 178
NEXN	GG - 33	GA - 6	AA - 0

1. NEXN
   (a) $P(Dados|GG) = 10^{-4.30} = 5.012 \times 10^{-5}$
   (b) $P(Dados|GT) = 10^{0} = 1$
   (c) $P(Dados|TT) = 10^{-17.80} = 1.585 \times 10^{-18}$

2. NEXN
   (a) $P(Dados|GG) = 10^{-3.30} = 5.012 \times 10^{-4}$
   (b) $P(Dados|GA) = 10^{-0.60} = 0.251$
   (c) $P(Dados|AA) = 10^{0} = 1$

No gene NEXN, para o primeiro caso é a heterozigotia da alteração com a referência, enquanto que no segundo caso é a homozigotia da alteração.

Para a alteração com cobertura 24 do gene NEXN, obtemos no programa *GeMS* um valor-p de 0.1213, sendo portanto rejeitada a hipótese de se tratar de um SNP para qualquer um dos níveis de significância usuais.

Na Tabela 3.25 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.25: Mutações detectadas no indivíduo 02192A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
CAV3	A por G	Intrão	97	59	Modificador
CAV3	A por G	Transcripto	97	59	Modificador
GLA	C por T	*downstream*	158	93	Modificador
LDB3	G por A	*upstream*	139	41	Modificador
LDB3	G por A	5'UTR	139	41	Modificador
LDB3	T por C	Intrão	138	49	Modificador
LDB3	T por C	*upstream*	138	49	Modificador
LDB3	C por T	Intrão	124	40	Modificador
LDB3	C por T	Transcripto	124	40	Modificador
LDB3	delecção de T	Intrão	81.5	73	Modificador
LDB3	delecção de T	Transcripto	81.5	73	Modificador
MYLK2	G por C	Intrão	143	39	Modificador
MYLK2	G por C	*upstream*	143	39	Modificador
MYLK2	T por G	Intrão	9.52	47	Modificador
MYLK2	T por G	*upstream*	9.52	47	Modificador
MYLK2	A por G	Intrão	159	89	Modificador
MYLK2	A por G	Transcripto	159	89	Modificador
PRKAG2	T por C	Intrão	177	91	Modificador
PRKAG2	T por C	Transcripto	177	91	Modificador
PRKAG2	T por C	*downstream*	177	91	Modificador

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
PRKAG2	G por A	Intrão	191	66	Modificador
PRKAG2	G por A	Transcripto	191	66	Modificador
PRKAG2	G por A	*downstream*	191	66	Modificador
PRKAG2	T por C	Intrão	131	71	Modificador
PRKAG2	T por C	Transcripto	131	71	Modificador
PRKAG2	T por C	*upstream*	131	71	Modificador
PRKAG2	T por C	*downstream*	131	71	Modificador
PRKAG2	A por C	Intrão	147	66	Modificador
PRKAG2	A por C	Transcripto	147	66	Modificador
PRKAG2	A por C	*upstream*	147	66	Modificador
PRKAG2	A por C	*downstream*	147	66	Modificador
PRKAG2	T por C	Intrão	99	40	Modificador
PRKAG2	T por C	Transcripto	99	40	Modificador
PRKAG2	T por C	*downstream*	99	40	Modificador
VCL	T por C	Intrão	181	47	Modificador
VCL	T por C	*downstream*	181	47	Modificador

Na Tabela 3.25 observa-se que o programa *snpEff*, com a base de dados *GRCh37.65*, apresenta variações altamente significativas em 6 genes não descritos em literatura.

Na Tabela 3.26 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.26: Valores-p das alterações detectadas no indivíduo 02192A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
CAV3	A por G	0.0000	Possível SNP
GLA	C por T	0.0000	Possível SNP
LDB3	G por A	0.0000	Possível SNP
LDB3	T por C	0.0000	Possível SNP
LDB3	C por T	0.0000	Possível SNP
MYLK2	G por C	0.0000	Possível SNP
MYLK2	T por G	0.0000	Possível SNP
MYLK2	A por G	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	G por A	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP

Gene	Alteração	Valor-p	Conclusão
PRKAG2	A por C	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP

Todas as alterações detectadas nos 6 genes são possíveis candidatos a SNP, de acordo com o programa *GeMS*.

3.8 Amostra 02193A

Na Tabela 3.27 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*. A alteração que apresenta valor para a cobertura não é significativa e para além disso são ambas comuns, ainda que nada se saiba da cobertura da última do gene NEXN.

Tabela 3.27: Mutações detectadas no indivíduo 02193A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
NEXN	C por T	Intrão	112	19	NA
NEXN	inserção de AAA	Intrão	18.5	NA	NA

Na Tabela 3.28 encontram-se os valores das verosimilhanças dos genótipos das alterações detectadas utilizando o programa *ANNOVAR* com a base de dados *hg19*, na escala PHRED.

Tabela 3.28: Verosimilhança do genótipo em escala PHRED das alterações detectadas no indivíduo 02193A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	Referência	Alteração e Referência	Alteração
NEXN	CC - 142	CT - 0	TT - 129
NEXN	Nenhum - 56	-A - 0	AA - 56

1. NEXN
   (a) $P(Dados|CC) = 10^{-14.20} = 6.310 \times 10^{-15}$
   (b) $P(Dados|CT) = 10^{0} = 1$

(c) $P(Dados|TT) = 10^{-12.90} = 1.259 \times 10^{-13}$

2. NEXN

(a) $P(Dados|--) = 10^{-5.60} = 2.512 \times 10^{-6}$
(b) $P(Dados|-A) = 10^{0} = 1$
(c) $P(Dados|AA) = 10^{-5.60} = 2.512 \times 10^{-6}$

No NEXN, para ambos os casos é a heterozigotia a situação mais provável.

Na Tabela 3.29 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.29: Mutações detectadas no indivíduo 02193A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
ANKRD1	delecção de TG	Intrão	145	93	Modificador
CAV3	G por T	Transcripto	89	57	Modificador
CAV3	G por T	*upstream*	89	57	Modificador
LDB3	T por C	Transcripto	154	74	Modificador
LDB3	T por C	*upstream*	154	74	Modificador
LDB3	A por T	Intrão	113	55	Modificador
LDB3	A por T	Transcripto	113	55	Modificador
MYLK2	T por C	Factor de iniciação ganho	109	47	Baixo
MYLK2	T por C	5'UTR	109	47	Modificador
MYLK2	G por C	Intrão	119	52	Modificador
MYLK2	G por C	*upstream*	119	52	Modificador
PRKAG2	G por A	Intrão	89	97	Modificador
PRKAG2	G por A	Transcripto	89	97	Modificador
PRKAG2	G por A	*downstream*	89	97	Modificador
PRKAG2	A por C	Intrão	131	88	Modificador
PRKAG2	A por C	Transcripto	131	88	Modificador
PRKAG2	A por C	*downstream*	131	88	Modificador
PRKAG2	A por C	*upstream*	131	88	Modificador
PRKAG2	T por C	Intrão	61	43	Modificador
PRKAG2	T por C	Transcripto	6 1	43	Modificador
PRKAG2	T por C	*downstream*	61	43	Modificador

Gene	Alteração	Local	Qualidade	Cobertura	Impacto
PRKAG2	A por G	Intrão	188	49	Modificador
PRKAG2	A por G	Transcripto	188	49	Modificador
VCL	T por C	Intrão	127	74	Modificador
VCL	T por C	*downstream*	127	74	Modificador
VCL	T por C	Intrão	106	52	Modificador
VCL	T por C	*downstream*	106	52	Modificador

Reccorendo ao programa *snpEff*, detectaram-se 6 genes não descritos na literatura com alterações significativas, conforme apresentado na Tabela 3.29.

Na Tabela 3.30 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.30: Valores-p das alterações detectadas no indivíduo 02193A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	Alteração	Valor-p	Conclusão
CAV3	G por T	0.0000	Possível SNP
LDB3	T por C	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
MYLK2	T por C	0.0000	Possível SNP
MYLK2	G por C	0.0000	Possível SNP
PRKAG2	G por A	0.0000	Possível SNP
PRKAG2	A por C	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP
VCL	T por C	0.0464	Possível SNP

Todas alterações detectadas pelo *snpEff* são consideradas candidatas a SNP pelo *GeMS*, ainda que no último caso, o valor-p esteja demasiado próximo de 0.05, é inferior.

3.9 Amostra 02194A

Na Tabela 3.31 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*. O programa *ANNOVAR* não detectou qualquer variação significativa, visto que possuem cobertura inferior a 30 e a variação de NEXN é sinónima.

Tabela 3.31: Mutações detectadas no indivíduo 02194A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
JPH2	C por T	Exão	NA	NA	Sinónima
MYBPC3	G por A	Exão	NA	NA	Sinónima
MYBPC3	C por T	Exão	225	78	Sinónima
NEXN	inserção de A	Intrão	34.2	NA	Modificador
NEXN	G por T	Intrão	31.8	2	Modificador
NEXN	C por T	Intrão	31.8	2	Modificador

Na Tabela 3.32 encontram-se os valores das verosimilhanças dos genótopos das alterações detectadas utilizando o programa *ANNOVAR* com a base de dados *hg19*, na escala PHRED.

Tabela 3.32: Verosimilhança do genótipo em escala PHRED das alterações detectadas no indivíduo 02194A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Referência**	**Alteração e Referência**	**Alteração**
NEXN	Nenhum - 73	-A - 6	AA - 0
NEXN	GG - 63	GT - 6	TT - 0
NEXN	CC - 63	CT - 6	TT - 0

1. NEXN
   (a) $P(Dados|--) = 10^{-7.30} = 5.012 \times 10^{-8}$
   (b) $P(Dados|-A) = 10^{-0.60} = 0.251$
   (c) $P(Dados|AA) = 10^{0} = 1$
2. NEXN
   (a) $P(Dados|GG) = 10^{-6.30} = 5.012 \times 10^{-7}$
   (b) $P(Dados|GT) = 10^{-0.60} = 0.251$
   (c) $P(Dados|TT) = 10^{0} = 1$
3. NEXN
   (a) $P(Dados|CC) = 10^{-6.30} = 5.012 \times 10^{-7}$
   (b) $P(Dados|CT) = 10^{-0.60} = 0.251$
   (c) $P(Dados|TT) = 10^{0} = 1$

Para o gene NEXN, para todas as alterações, o genótipo mais provável é a homozigotia da alteração.

De acordo com o programa GeMS, as susbtituições do gene NEXN possuem um valor-p de 0.0000, sendo consideradas como possíveis candidatas a SNP.

Na Tabela 3.33 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.33: Mutações detectadas no indivíduo 02194A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
ANKRD1	delecção de TG	Intrão	199	96	Modificador
CAV3	A por G	Intrão	147	98	Modificador
CAV3	A por G	Transcripto	147	98	Modificador
LDB3	G por A	Intrão	127	59	Modificador
LDB3	G por A	5'UTR	127	59	Modificador
LDB3	G por A	*upstream*	127	59	Modificador
LDB3	A por G	Intrão	110	62	Modificador
LDB3	A por G	*upstream*	110	62	Modificador
LDB3	A por G	5'UTR	110	62	Modificador
LDB3	T por C	Intrão	130	75	Modificador
LDB3	T por C	*upstream*	130	75	Modificador
LDB3	A por G	Intrão	143	44	Modificador
LDB3	A por G	Transcripto	143	44	Modificador
LDB3	A por T	Intrão	115	48	Modificador
LDB3	A por T	Transcripto	115	48	Modificador
LDB3	G por A	Intrão	121	32	Modificador
LDB3	G por A	Transcripto	121	32	Modificador
MYLK2	T por C	Factor de iniciação ganho	137	45	Baixo
MYLK2	T por C	5'UTR	137	45	Modificador
MYLK2	G por C	Intrão	195	70	Modificador
MYLK2	G por C	*upstream*	195	70	Modificador
MYOZ2	T por C	Intrão	125	53	Modificador
PRKAG2	A por T	Intrão	136	33	Modificador
PRKAG2	A por T	Transcripto	136	33	Modificador

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
PRKAG2	A por T	*downstream*	136	33	Modificador
PRKAG2	A por T	*upstream*	136	33	Modificador
PRKAG2	G por A	Intrão	115	67	Modificador
PRKAG2	G por A	Transcripto	115	67	Modificador
PRKAG2	G por A	*downstream*	115	67	Modificador
PRKAG2	T por C	Intrão	141	73	Modificador
PRKAG2	T por C	Transcripto	141	73	Modificador
PRKAG2	T por C	*upstream*	141	73	Modificador
PRKAG2	A por G	Intrão	124	33	Modificador
PRKAG2	A por G	Transcripto	124	33	Modificador
VCL	T por C	Intrão	149	86	Modificador
VCL	T por C	*downstream*	149	86	Modificador

Com o programa *snpEff* foram detectados 7 genes não descritos na literatura, todos com alterações significativas (Tabela 3.33).

Na Tabela 3.34 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.34: Valores-p das alterações detectadas no indivíduo 02194A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
CAV3	A por G	0.0000	Possível SNP
LDB3	G por A	0.0000	Possível SNP
LDB3	A por G	0.0000	Possível SNP
LDB3	T por C	0.0000	Possível SNP
LDB3	A por G	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
LDB3	G por A	0.0000	Possível SNP
MYLK2	T por C	0.0000	Possível SNP
MYLK2	G por C	0.0000	Possível SNP
MYOZ2	T por C	0.0000	Possível SNP
PRKAG2	A por T	0.0000	Possível SNP
PRKAG2	G por A	0.0000	Possível SNP
PRKAG2	T por C	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP

Todas as substituições foram consideradas como possíveis SNP pelo *GeMS*.

3.10 Amostra 02195A

Na Tabela 3.35 encontram-se descritas as mutações identificadas pelo programa *ANNOVAR* com a base de dados *hg19*. O programa *ANNOVAR* apenas fornece uma substituição no gene MYBPC3 significativa, mas trata-se de uma mutação sinónima e comum.

Tabela 3.35: Mutações detectadas no indivíduo 02195A, utilizando o programa *ANNOVAR* com a base de dados *hg19*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
MYBPC3	G por A	Exão	225	85	Sinónima

Na Tabela 3.36 encontram-se descritas as mutações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*.

Tabela 3.36: Mutações detectadas no indivíduo 02195A, utilizando o programa *snpEff* com a base de dados *GRCh37.65*.

Gene	**Alteração**	**Local**	**Qualidade**	**Cobertura**	**Impacto**
CAV3	C por G	Intrão	9.52	30	Modificador
CAV3	C por G	Transcripto	9.52	30	Modificador
LDB3	T por C	Intrão	100	75	Modificador
LDB3	T por C	*upstream*	100	75	Modificador
LDB3	T por C	*downstream*	100	75	Modificador
LDB3	A por T	Intrão	131	48	Modificador
LDB3	A por T	Transcripto	131	48	Modificador
MYLK2	G por C	Intrão	28	50	Modificador
MYLK2	G por C	*upstream*	28	50	Modificador
MYOZ2	C por T	3'UTR	97	56	Modificador
PRKAG2	A por G	Intrão	138	36	Modificador
PRKAG2	A por G	Transcripto	138	36	Modificador
VCL	T por C	Intrão	112	85	Modificador
VCL	T por C	*downstream*	112	85	Modificador

Foram detectados 6 genes pelo programa *snpEff*, todos com alterações significativas, conforme descrito na Tabela 3.36.

Na Tabela 3.37 encontram-se os valores-p das alterações identificadas pelo programa *snpEff* com a base de dados *GRCh37.65*, com recurso ao programa *GeMS*.

Tabela 3.37: Valores-p das alterações detectadas no indivíduo 02195A - *snpEff* - *GRCh37.65* no programa *GeMS*.

Gene	**Alteração**	**Valor-p**	**Conclusão**
CAV3	C por G	0.0000	Possível SNP
LDB3	T por C	0.0000	Possível SNP
LDB3	A por T	0.0000	Possível SNP
MYLK2	G por C	0.0000	Possível SNP
MYOZ2	C por T	0.0000	Possível SNP
PRKAG2	A por G	0.0000	Possível SNP
VCL	T por C	0.0000	Possível SNP

Todas as substituições foram detectadas como possíveis SNP pelo programa GeMS.

3.11 Conclusões

Dos oito genes não descritos na literatura (ANKRD1, CAV3, GLA, LDB3, MYLK2, MYOZ2, PRKAG2 e VCL), apresentam-se os seus alinhamentos ao genoma de referência com os indivídous que apresentam alterações significativas para esses genes. Três dos genes detectados, apareceram em todos os indíviduos (LDB3, PRKAG2 e VCL).

O programa *snpEff* aparenta ser mais sensível que o *ANNOVAR*, pois foi o único que detectou as variações significativas. Quanto às bases de dados, aparentemente os resultados obtidos utilizando uma ou outra são muito semelhantes, com a diferença que a *GRCh37.65* detecta mais efeitos.

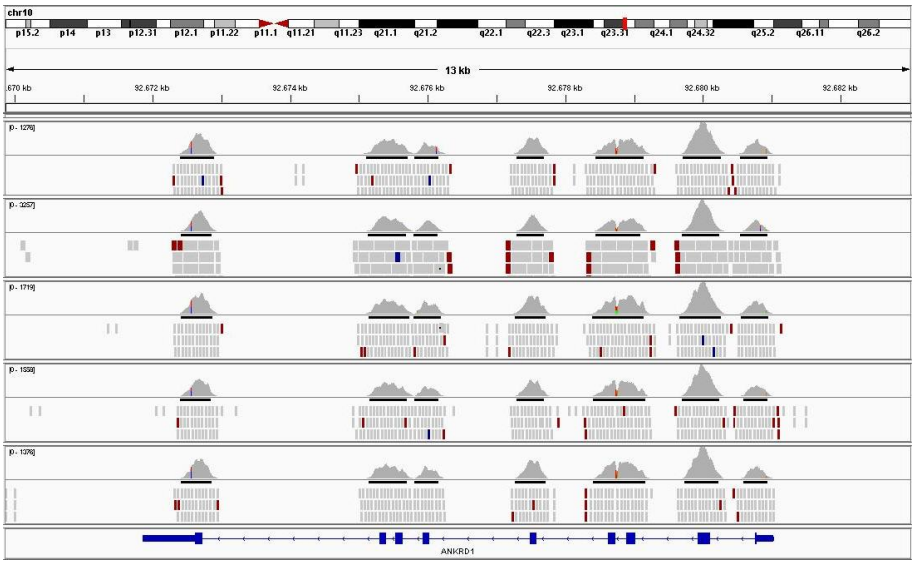

Figura 3.2: Alinhamento relativo ao gene ANKRD1, para os indivíduos 02186, 02187, 02189, 02193 e 02194 no programa *IGV*

O gene ANKRD1 codifica o domínio repetitivo da proteína anquinina. Um trabalho de Arimura e sua equipa, descreve três mutações exónicas associadas à patologia e quatro intrónicas. [Arimura *et al.*, 2009]

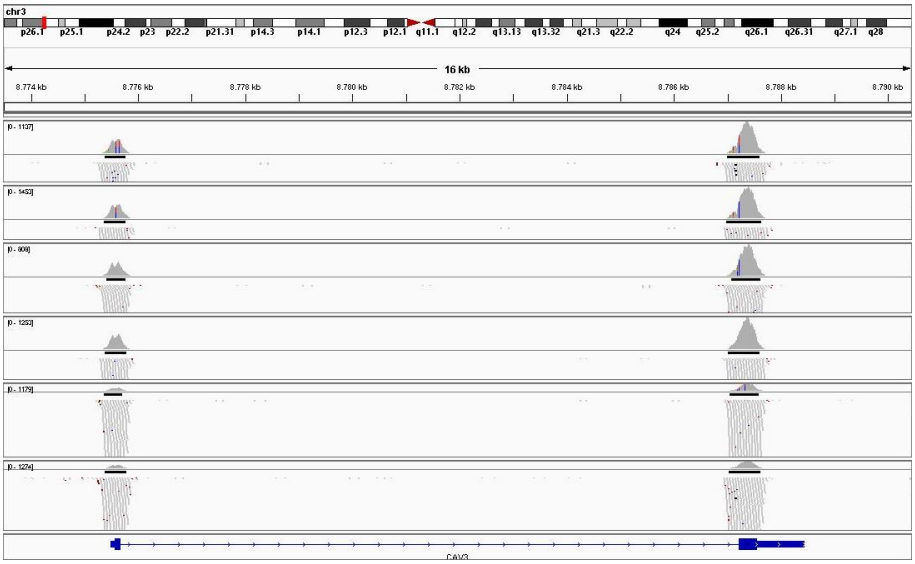

Figura 3.3: Alinhamento relativo ao gene CAV3, para os indivíduos 02186, 02188, 02192, 02193, 02194 e 02195 no programa *IGV*

O gene CAV3 codifica a caveolina-3, presente no miócito, e foi descrito por Hayashi e Gazzerro e suas equipas como sendo associado à doença por duas mutações a nível exónico, no nosso projecto foram detectadas alterações intrónicas. [Gazzerro *et al.*, 2010] [Hayashi *et al.*, 2004]

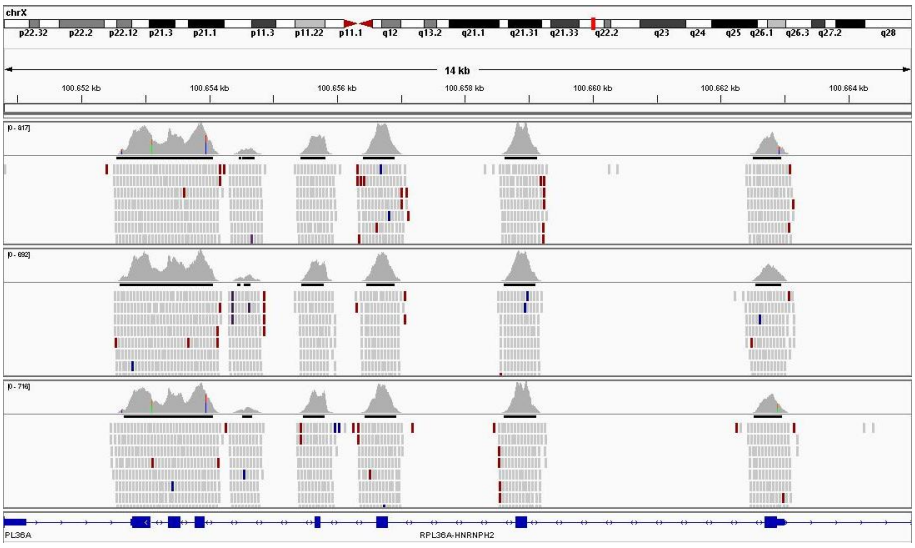

Figura 3.4: Alinhamento relativo ao gene GLA, para os indivíduos 02186, 02188 e 02192 no programa *IGV*

GLA codifica a α-galactosidase A e apenas é descrita por Monserrat e a sua equipa como associada à Doença de Fabry com uma mutação a nível exónico, sendo que associam esta patologia à Miocardiopatia Hipertrófica em 1% dos casos. [Monserrat *et al.*, 2007]

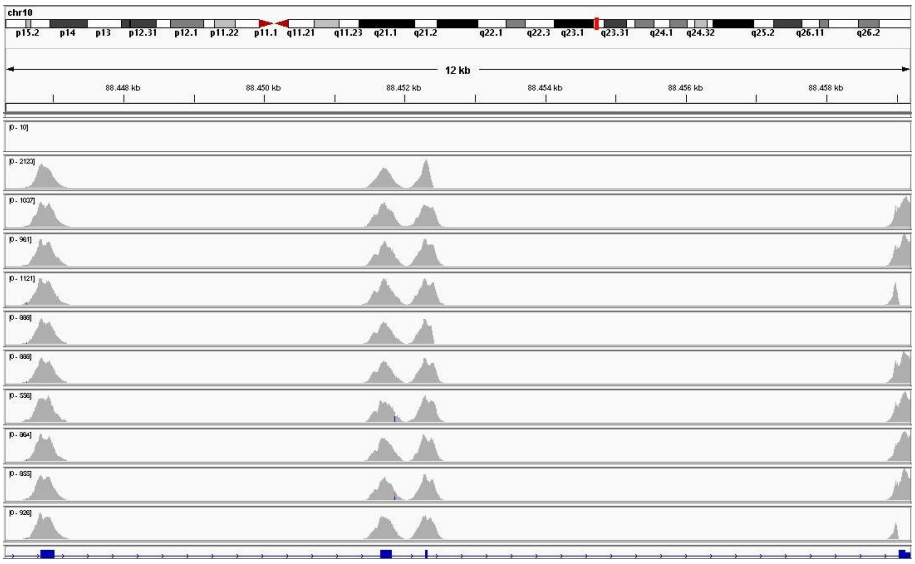

Figura 3.5: Alinhamento relativo ao gene LDB33, para os indivíduos 02186, 02187, 02188, 02189, 02190, 02191, 02192, 02193, 02194 e 02195 no programa *IGV*

O gene LDB3 é responsável pelo adaptador no músculo estriado para acoplar a proteína-C cinase. Só foram descritas mutações a nível exónico associadas à doença em estudo. [Kimura, 2010] [Theis *et al.*, 2006]

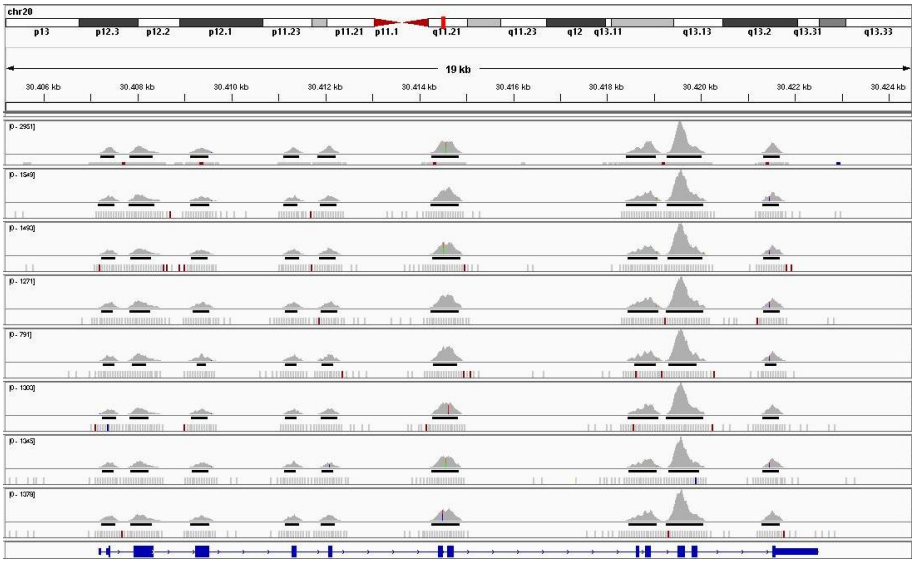

Figura 3.6: Alinhamento relativo ao gene MYLK2, para os indivíduos 02187, 02188, 02189, 02191, 02192, 02193, 02194 e 02195 no programa *IGV*

O gene MYLK2 codifica um enzima cinase da cadeia leve de miosina. Apenas foi descrita uma mutação exónica associada à Miocardiopatia Hipertrófica. [Brion *et al.*, 2008]

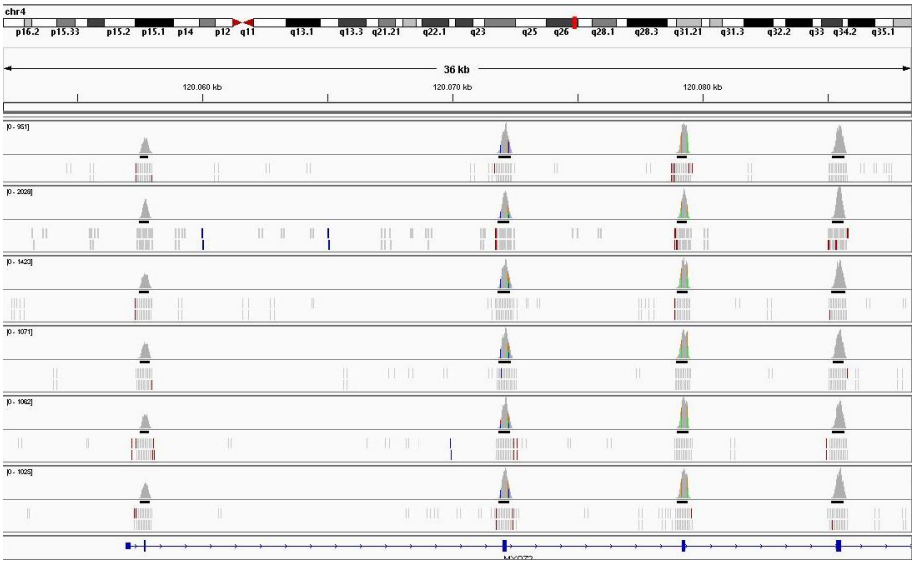

Figura 3.7: Alinhamento relativo ao gene MYOZ2, para os indivíduos 02186, 02187, 02189, 02191, 02194 e 02195 no programa *IGV*

O gene MYOZ2 codifica a miozenina 2 que é uma proteína do disco-Z. Osio e sua

equipa descreve uma mutação exónica associada à patologia em estudo e uma revisão de 2010 considera um gene já totalmente relacionado com a doença de forma estabelecida. [Marian, 2010] [Osio *et al.*, 2007]

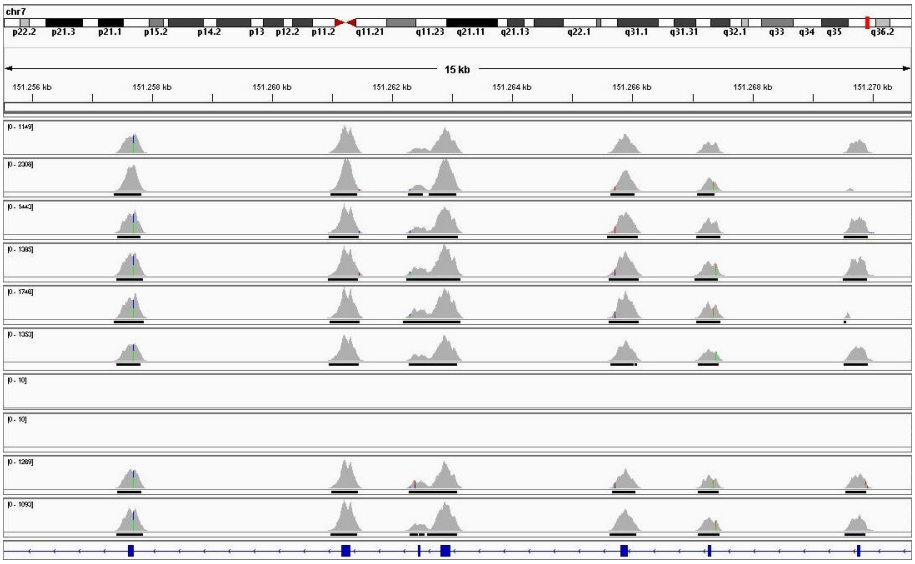

Figura 3.8: Alinhamento relativo ao gene PRKAG2, para os indivíduos 02186, 02187, 02188, 02189, 02190, 02191, 02192, 02193, 02194 e 02195 no programa *IGV*

O gene PRKAG2 codifica o enzima cinase activado por AMP. Foi associado à doença a nível exónico por Fokstuen e sua equipa. [Fokstuen *et al.*, 2008]

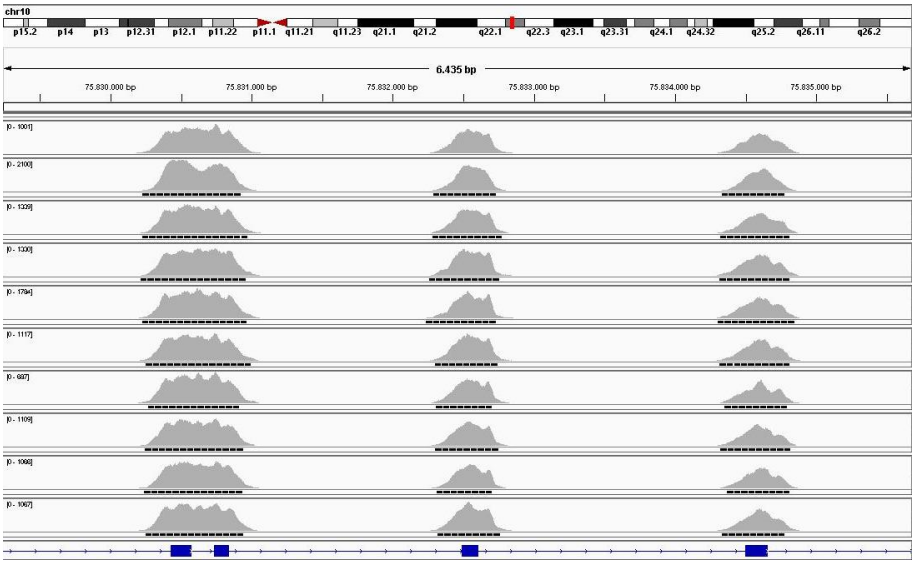

Figura 3.9: Alinhamento relativo ao gene VCL, para os indivíduos 02186, 02187, 02188, 02189, 02190, 02191, 02192, 02193, 02194 e 02195 no programa *IGV*

O gene VCL codifica a vinculina e está associado à doença por três mutações exónicas. [Theis *et al.*, 2006]

3.12 Discussão

Com apenas dez indivíduos, sequenciados uma vez, torna-se complicado retirar conclusões definitas relativamente a que genes desconhecidos poderão estar associados à Miocardiopatia Hipertrófica. São necessários mais indivíduos e sequenciações de modo a obter relações mais concretas, de qualquer modo, encontraram-se indícios fortes de novos genes que poderão, no futuro, ser associados à patologia.

Diversas poderiam ser as escolhas em relação aos programas a utilizar para qualquer uma das partes da metodologia. O procedimento apresentado trata-se do final, após o teste de diferentes *softwares*. Para o alinhamento experimentou-se o programa *Bowtie*, bastante semelhante ao *BWA*, e o programa *Maq*. Este último não foi escolhido pela última actualização ser do ano 2008 e pelas dificuldades encontradas no seu uso. [Langmead *et al.*, 2009]

Na fase de determinação de SNP, testou-se o programa *FreeBayes* que revelou ser bastante próximo do *SAMtools*. Dada a semelhança entre os resultados, optou-se pelo *SAMtools*, por ser o mais utilizado pelos utilizadores do fórum **SEQanswers** (*the next generation sequencing community* - http:\\seqanswers.com). Outro programa testado foi o *GATK*. Este programa aparece na metodologia logo após o alinhamento realizado. Como

encontrei diversos problemas ao utilizá-lo, acabei por desistir por conselho do investigador Tobias Rausch do GeneCore do EMBL (European Molecular Biology Laboratory) no grupo Korbel, dado que mesmo na sua equipa estavam com dificuldades em utilizá-lo. Seguindo o seu conselho optei pelo *ANNOVAR* e *snpEff*. [Li *et al.*, 2012]

Um dos problemas encontrados neste trabalho foi a falta de manuais de instruções completos dos programas a utilizar para a análise dos dados, sendo que muitos dos comandos não eram de todo totalmente descritos nestes manuais, sendo que muitas instruções só são descobertas em fóruns da área, como o **SEQanswers**. Outro desafio prende-se com a pouca informação disponibilizada sobre o funcionamento dos programas. Estas questões prendem-se com a novidade da técnica e a falta de metodologias padrão definidas. Os artigos encontrados não são completos em relação às descrições das metodologias estatísticas. Por outro lado, os programas funcionam como caixas negras, não sendo possível perceber muito bem como os programas implementam as metodologias.

Neste projecto tornam-se claras as vantagens desta nova metodologia, pois trazendo mais rapidez e permitindo uma análise a um maior número de genes, com custos reduzidos, permite que se analizem todos os genes que já se sabem estar associados à doença, ao contrário do que se faz actualmente em termos de diagnóstico. Por exemplo, neste momento, o número máximo de genes a que se faz rastreio é dez, sendo que muito raramente são pedidos todos.

A Sequenciação de Nova Geração oferece novas e melhores possibilidades no diagnóstico clínico, começando já a ser amplamente usada na área da investigação, não só nas Ciências da Saúde, como em Biologia.

Apêndice A

Comandos utilizados na análise

Para a análise dos dados relativos à Miocardiopatia Hipertrófica, determinou-se a seguinte ordem de instruções na linha de comandos do Linux:

1. **Alinhamento das amostras com o genoma de referência**

   i. `bwa aln -t 8 hg19RefGenome.fa sample_1.fastq > sample_1_align.sai`

   ii. `bwa aln -t 8 hg19RefGenome.fa sample_2.fastq > sample_2_align.sai`

   iii. `bwa sampe hg19RefGenome.fa sample_1_align.sai sample_2_align.sai sample_1.fastq sample_2.fastq > sample_align.sam -r "@RG\tID:sample\tLB:sample\tPL:ILLUMINA\tSM:sample"`

   O comando *aln* vai dizer ao programa para efectuar o primeiro passo do alinhamento para cada ficheiro. O parâmetro -t refere-se ao número de processadores utilizados, neste caso 8. É necessário realizar o primeiro passo do alinhamento aos dois ficheiros *.fastq*, pois tratam-se de leituras *paired-end*, também por esta razão se utiliza o comando *sampe* para a segunda parte do alinhamento. Os ficheiros *.sai* possuem as coordenadas das leituras, em relação ao genoma de referência calculadas pelo *BWA* (como explicado na secção 2.3, na página 11). O parâmetro -r serve para indicar a máquina utilizada para a sequenciação, neste caso *Illumina*.

2. **Converter a *.bam***

   i. `samtools view -bS sample_align.sam -o sample_align.bam -t hg19RefGenome.fa`

   O comando *view* vai converter o ficheiro *.sam* a *.bam*. O parâmetro - b indica que queremos o *output* no format *.bam*, -S que o *input* encontra-se no formato *.sam*, -o indica qual o ficheiro que será gerado como *output* e por último -t qual o ficheiro de referência.

3. **Organizar o ficheiro *.bam***

   i. `samtools sort sample_align.bam sample_align_sorted`
   O comando *sort* irá organizar o ficheiro *.bam* e criará um novo ficheiro já organizado.

4. **Indexação do ficheiro *.bam***

   i. `samtools index sample_align_sorted.bam`
   O comando *index* vai permitr ao programa ter acesso remoto ao ficheiro *.bam*, para concretizar as suas tarefas com mais eficiência.

5. **Determinar SNP e inserções/delecções**

   i. `samtools mpileup -P ILLUMINA -d8000 -B -uf hg19RefGenome.fa sample_align_sorted.bam | bcftools view -bcg - > varSample.bcf`

   ii. `bcftools view varSample.bcf | vcfutils.pl varFilter -D 100 > varSample.flt.vcf`
   O comando *mpileup* vai gerar um ficheiro *.bcf* a partir do ficheiro *.bam* indexado e organizado. O parâmetro -P serve para indicar que tecnologia foi utilizada na sequenciação, neste caso *Illumina*, o -d indica o número máximo de leituras para cada posição (8000), o -B ajuda a reduzir os falsos positivos que surgem quando há erros no alinhamento, ao desactivar a opção de realinhamento tendo em conta a qualidade de mapeamento por base, -u indica que queremos o *output* não comprimido e -f que o genoma de referência está no formato *.fasta*. Utilizando o símbolo |, mandamos a linha de comandos realizar outra tarefa assim que terminar a anterior com sucesso. Neste caso vamos querer correr o `bcftools` que é uma ferramenta do *SAMtools*. Com o comando *view* vamos converter o *output* anterior (no formato *.pileup*) num *.bcf*. O parâmetro -b indica que queremos um output *.bcf*, -c que estamos a determinar variantes recorrendo a inferência bayesiana e -g para reforçar o -c. Na segunda parte deste processo, vamos converter o *.bcf* a *.vcf* com o comando *view* e, usando o *output* gerado, vamos filtrá-lo usando o *script vcfutils.pl* do `bcftools` com recurso ao parâmetro -D que limita a cobertura máxima por leitura (deve ser o dobro da cobertura média real, que no nosso caso é 50).

6. **Anotação dos SNP**

   (a) Utilizando snpEFF:

      i. `java -Xmx4G -jar snpEff_2_0_5d/snpEff.jar eff -c snpEff_2_0_5d/snpEff.config -v -onlyCoding true -i vcf -o txt`

`hg19 varSample.flt.vcf > snpEff_sample.txt`

O programa *snpEff* necessita do comando *java*, pois foi programado nessa linguagem. Ao indicarmos *-Xmx4G* queremos que o processador utilize 4 Gb de memória RAM. Utilizando o *script snpEff_2_0_5d/snpEff.jar*, vamos especificar que determine os efeitos das variações com o comando *eff*. Com -c especificamos o ficheiro de configuração (*snpEff_2_0_5d/snpEff.config*), com -v estamos a pedir que o programa nos explique o que está a fazer, com -onlyCoding na versão *true* estamos a pedir apenas as regiões codificantes, -i indica que o *input* é um *.vcf* e -o que o *output* será um *.txt*. O *hg19* indica a base de dados de variação a utilizar.

(b) Utilizando ANNOVAR:

i. `convert2annovar.pl --format vcf4 --includeinfo varSample.flt.vcf > sample.snp.annovar`

ii. `summarize_annovar.pl -builver hg19 -remove -verdbsnp 132 -ver1000g 1000g2012feb sample.snp.annovar -outfile sample.variant humandb`

Usando o *script convert2annovar.pl* do programa *ANNOVAR*, vamos converter o ficheiro *.vcf* (indicado pelo parâmetro –format *vcf4*) ao tipo de ficheiro usado pelo programa. A opção –includeinfo permite-nos não perder nenhuma da informação contida no *.vcf* original. Com o *script summarize_annovar.pl*, vamos anotar as variações. Com -builver indicamos a base de dados (*hg19*), com -remove pedimos que o programa não escreva na linha de comandos o *output*, com -verdbsnp indicamos a versão da base de dados dos SNP (*132*), com -ver1000g a versão do projecto 1000G a utilizar e, por último, com -outfile com o *output*. O parâmetro *humandb* indica apenas a pasta onde se encontram as bases de dados.

Referências

[Ansorge, 2009] Ansorge, W.J., "Next-generation DNA sequencing techniques," *New Biotechnology*, Vol. 25, No. 4, pp. 195–203, Abril 2009.

[Arimura *et al.*, 2009] Arimura, T., Bos, J.M., Sato, A., Kubo, T., Okamoto, H., Nishi, H., Harada, H., Koga, Y., Moulik, M., Doi, Y.L., Towbin, J.A., Ackerman, M.J. e Kimura, A., "Cardiac Ankyrin Repeat Protein Gene (ANKRD1) Mutations in Hypertrophic Cardiomyopathy," *Journal of the American College of Cardiology*, Vol. 54, No. 4, pp. 334–342, Julho 2009.

[Brion *et al.*, 2008] Brion, M., Allegue, C., Monserrat, L., Hermida, M., Castro-Beiras, A., Carracedo, A., "Large scale analysis of HCM mutations in sudden cardiac death, " *Forensic Sciences International: Genetics Supplement*, Série I, pp. 549–550, 2008.

[Cingolani, 2012] Cingolani, P., "snpEff: Variant effect prediction,"http://snpeff.sourceforge.net, 2012.

[Cirino e Ho, 2011] Cirino, A.L. e Ho, C., "Familial Hypertrophic Cardiomyopathy Overview," *GeneReviews [Internet]*, Agosto 2008 (Actualizado a Maio 2011).

[Cock *et al.*, 2010] Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. e Rice, P.M., "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acid Research*, Vol 38, No. 6, pp. 1767–1771, 2010.

[Danecek *et al.*, 2011] Danecek, P., Auton, A., Abecassis, G., Albers, C.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R. e 1000 Genomes Project Analysis Group, "The variant call format and VCFtools," *Bioinformatics*, Vol. 27, No. 15, pp. 2156–2158, Junho 2011.

[Deorowicz e Grabowski, 2011] Deorowicz, S. e Grabowski, S., "Compression of DNA sequence reads in FASTQ format," *Bioinformatics*, Vol. 27, No. 6, pp. 860–862, Janeiro 2011.

[Dixon, 1950] Dixon, W.J., "Analysis of extreme values," *The Annals of Mathematical Statistics*, Vol. 21, No. 4, pp. 488–506, Dezembro 1950.

[Flicek *et al.*, 2011] Flicek, P. *et al.*, "Ensembl 2012,", *Nucleic Acids Research*, Vol. 40, *Database Issue*, 2012.

[Fokstuen *et al.*, 2008] Fokstuen, S., *et al.*, "A DNA Resequencing Array for Pathogenic Mutation Detection in Hypertrophic Cardiomyopathy," *Human Mutation*, Vol. 29, No. 6, pp. 879–885, 2008.

[Fujita *et al.*, 2011] Fujita, P.A. *et al.*, "The UCSC Genome Browser database: update 2011," *Nucleic Acids Research*, Vol. 39, *Database Issue*, 2011.

[Gazzerro *et al.*, 2010] Gazzerro, E., Sotgia, F., Bruno, C., Lisanti, M.P. e Minetti, C., "Caveolinopathies: from the biology of caveolin-3 to human diseases," *European Journal of Human Genetics*, Vol. 18, pp. 137–145, 2010.

[Hartwell *et al.*, 2008] Hartwell, L.H., Hood, L., Goldber, M.L., Reynolds, A.E., Silver, L.M. e Veres, R.C., "Genetics: From Genes to Genome, 3^{rd} Edition," *Mc Graw Hill*, 2008.

[Hayashi *et al.*, 2004] Hayashi, T., *et al*, "Identification and functional Analysis of a caveolin-3 mutation with familial hypertrophic cardiomyopathy," *Biochemical and Biophysical Research Comunications*, Vol. 313, pp. 178–184, 2004.

[Ho, 2011] Ho, C.,"Hypertrophic Cardiomyopathy: *For Heart Failure Clinics*: Genetics of Cardiomyopathy and Heart Failure," *Heart Failure Clinics*, Vol. 6, No. 2, pp. 141–159, Abril 2010.

[Junqueira e Carneiro, 2003] Junqueira, L.C. e Carneiro, J., "Basic Histology, 10^{th} Edition," *Lange Medical Books Mc Graw Hill*, 2003.

[Kimura, 2010] Kimura, A., "Molecular basis of hereditary cardiomyopathy: abnormalities in calcium sensitivity, stretch response, stress response and beyond," *Journal of Human Genetics*, Vol. 55, pp. 81–90, 2010.

[Langmead *et al.*, 2009] Langmead, B., Trapnell, C., Pop M., e Salzberg, S.L., "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, Vol. 10, No. 3, Artigo R25, Março 2009.

[Li, 2010] Li, H., "Mathematical Notes on SAMtools Algorithms,"http://bit.ly/stmath, Outubro 2010.

[Li, 2011] Li, H., "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics*, Vol. 27, No, 21, pp. 2987–2993, Setembro 2011.

[Li e Durbin, 2009] Li, H. e Durbin, R., "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, Vol. 25, No. 14, pp. 1754–1760, Maio 2009.

[Li *et al.*, 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecassis, G., Durbin, R., e 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, Vol.25, No. 16, pp. 2078–2079, Junho 2009.

[Li *et al.*, 2012] Li, J., Schmieder, R., Ward, R.M., Delenick, J., Olivares, E.C e Mittelman, D., "SEQanswers: an open access community for collaboratively decoding genomes," *Bioinformatics*, Vol. 28, No. 9, pp. 1272–1273, Março 2012.

[Lindblom e Robinson, 2011] Lindblom, A. e Robinson, P.N.,"Bioinformatics for Human Genetics: Promises and Challenges," *Human Mutation*, Vol. 32, No. 5, pp. 495–500, Fevereiro 2011.

[Lodish et al., 2008] Lodish, H., Berk, A., Kaiser, C.A., Krieger, M., Scott, M.P., Bretscher, A., Ploegh, H. e Matsudaira, P., "Molecular Cell Biology, 6^{th} Edition," *W.H. Freeman and Company*, 2008.

[Magi *et al.*, 2010] Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., e Brandi, M.L., "Bioinformatics for Next Generation Sequencing Data," *Genes*, Vol. 1, No. 2, pp. 294–307, Setembro 2010.

[Marian, 2010] Marian, A.J., "Hypertrophic cardiomyopathy: from genetics to treatment," *European Journal of Clinical Investigation*, Vol. 40, No. 4, pp. 360–69, 2010.

[McCahill, 1996] McCahill, T.A., "Factos Básicos em Biologia, 4^a Edição," *Editora Replicação*, 1996.

[Monserrat *et al.*, 2007] Monserrat, L., *et al.*, "Prevalence of Fabry Disease in a Cohort of 508 unrelated Patients With Hypertrophic Cardiomyopathy," *Journal of the American College of Cardiology*, Vol.50, No. 25, pp. 2399–2403, Dezembro 2007.

[Nielsen *et al.*, 2011] Nielsen, R., Paul, J.S., Albrechtsen, A. e Song, Y.S., "Genotype and SNP calling from next-generation sequencing data," *Nature Reviews Genetics*, Vol. 12, pp. 443–451, Junho 2011.

[Nowrousian, 2010] Nowrousian, M., "Next-Generation Sequencing Techniques for Eukaryotic Microorganisms: Sequencing-Based Solutions to Biological Problems," *Eukaryotic Cell*, Vol. 9, No. 9, pp. 1300–1310, Setembro 2010.

[Osio *et al.*, 2007] Osio, A., Tan, L., Chen, S.N., Lombardi, R., Nagueh, S.F., Shete, S., Roberts, R., Willerson, J.T., e Marian, A.J., "Myozenin 2 Is a Novel Gene for Human Hypertrophic Cardiomyopathy,", *Circulation Research*, Vol. 100, pp.766-768, 2007.

[Robinson *et al.*, 2011] Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. e Mesirov, J.P., "Integrative Genomics Viewer," *Nature Biotechnology*, Vol. 29, No. 1, pp.24–26, Janeiro 2011.

[Seidman e Seidman, 2011] Seidman, C. E. e Seidman, J. G.,"Identifying Sarcomere Gene Mutations in Hypertrophic Cardiomyopathy: A Personal History," *Circulation Research*, No. 108, pp. 743–750, Abril 2011.

[Theis *et al.*, 2006] Theis, J.L., *et al.*, "Echocardiographic-determined septal morphology in Z-disc Hypertrophic Cardiomyopathy," *Biochemical and Biophysical Research Comminications*, Vol. 351, pp. 896–902, 2006.

[Thorvaldsdóttir, Robinson e Mesirov, 2012] Thorvaldsdóttir, H., Robinson, J.T. e Mesirov, J.P., "Integrative Genomic Viewer (IGV): high-performance genomic data visualization and exploration," *Briefings on Bioinformatics*, Abril 2012.

[Voelkerding, Dames e Durtschi, 2009] Voelkerding, K.V., Dames, S.A. e Durtschi, J.D., "Next-Generation Sequencing: From Basic Research to Diagnostics," *Clinical Chemistry*, Vol. 55, No. 4, pp. 641–658, Abril 2009.

[Voelkerding *et al.*, 2010] Voelkerding, K. V., Dames, S. e Durtschi, J.,"Next Generation Sequencing for Clinical Diagnostics-Principles and Application to Targeted Resequencing for Hypertrophic Cardiomyopathy," *Journal of Molecular Diagnostics*, Vol. 12, No. 5, pp. 539–551, Abril 2010.

[Wang *et al.*, 2010] Wang, H. *et al.*, "Mutations in *NEXN*, a Z-Disc gene, Are Associated with Hypertrophic Myocardiopathy," *The American Journal of Human Genetics*, Vol. 87, pp. 687-693, Novembro 2010.

[Wang, Li e Hakonarson, 2010] Wang, K., Li, M., Hakonarson, H., "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Research*, Vol. 38, No. 16, e164, Julho 2010.

[You *et al.*, 2012] You, N., Murillo, G., Su, X., Zeng, X., Xu, J., Ning, K., Zhang, S., Zhu, J. e Cui, X., "SNP calling using genotype model selection on high-troughput sequencing data," *Bioinformatics*, Vol. 28, No. 5, pp. 643–650, Julho 2012.

[Zhang *et al.*, 2011] Zhang, J., Chiodini, R., Badr, A. e Zhang, G.,"The impact of next-generation sequencing on genomics," *Journal of Genetics and Genomics*, Vol. 38, pp. 95–109, Fevereiro 2011.