ANNA FLAVIA FIGUEREDO BENEDETTI

Análise de diferentes métodos de sequenciamento de larga
escala dos genes envolvidos no hipopituitarismo e
embriogênese hipofisária

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

Programa de Endocrinologia

Orientador: Prof^a. Dr^a. Luciani Renata Silveira de Carvalho

Co-Orientador: Dr^a. Ana Carolina Tahira

São Paulo

2018

Dados Internacionais de Catalogação na Publicação (CIP)

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

©reprodução autorizada pelo autor

Benedetti, Anna Flávia Figueiredo

Análise de diferentes métodos de sequenciamento
de larga escala dos genes envolvidos no
hipopituitarismo e embriogênese hipofisária / Anna
Flávia Figueiredo Benedetti. -- São Paulo, 2018.

Dissertação (mestrado)--Faculdade de Medicina da
Universidade de São Paulo.

Programa de Endocrinologia.

Orientadora: Luciani Renata Silveira de Carvalho.

Coorientadora: Ana Carolina Tahira.

Descritores: 1.Hipopituitarismo 2.Biologia
computacional 3.Sequenciamento completo do exoma
4.Exoma 5.Kits de reagente para diagnóstico
6.Biologia molecular

USP/FM/DBD-525/18

Responsável: Erinalva da Conceição Batista, CRB-8 6755

Este trabalho foi desenvolvido na
Unidade de Endocrinologia do
Desenvolvimento, Laboratório de
Hormônios e Genética Molecular LIM/42
da Disciplina de Endocrinologia do
Hospital das Clínicas da Faculdade de
Medicina da Universidade de São Paulo

Normatização Adotada

Esta Tese está de acordo com as seguintes normas, em vigor no momento desta publicação:

Referências: adaptado de International Committee of Medical Journals Editors (Vancouver).

Universidade de São Paulo. Faculdade de Medicina. Divisão de Biblioteca e Documentação. Guia de apresentação de dissertações, teses e monografias. Elaborado por Anneliese Carneiro da Cunha, Maria Julia de A. L. Freddi, Maria F. Crestana, Marinalva de Souza Aragão, Suely Campos Cardoso, Valéria Vilhena. 3a ed. São Paulo: Divisão de Documentação; 2011.

Abreviaturas dos títulos dos periódicos de acordo com List of Journals Indexed in Index Medicus.

Sumário

Lista de abreviaturas

Lista de tabelas

Lista de figuras

Resumo

Summary

1	Introdução.....	1
1.1	Embriogênese hipofisária.....	1
1.2	Hipopituitarismo congênito.....	2
1.3	Sequenciamento por metodologia de Sanger.....	4
1.4	Sequenciamento de nova geração.....	4
2	Objetivos.....	12
2.1	Primário.....	12
2.2	Secundários.....	12
2.3	Considerações éticas.....	12
3	Metodologia.....	13
3.1	Pesquisa bibliográfica de genes a serem incluídos no estudo.....	13
3.2	Amostras.....	15
3.2.1	HapMap	15
3.2.2	Paciente com hipopituitarismo e sua mãe	16
3.2.3	Amostras aleatórias	17
3.3	Extração de DNA.....	18
3.4	Preparação das bibliotecas.....	20
3.4.1	Nextera.....	20
3.4.2	NimbleGen.....	20
3.4.3	SureSelect/SureSelect + UTR	21
3.5	Análise bioinformática.....	22
3.6	ABraOM.....	24
4.	Resultados.....	26
4.1	Qualidade do sequenciamento.....	28
4.2	Qualidade do mapeamento e taxa de duplicação.....	30
4.3	Profundidade de cobertura global.....	32
4.4	Probes desenhadas para a região de interesse.....	36

4.5 Profundidade de cobertura na região alvo.....	37
4.6 Cobertura das regiões de acordo com a quantidade de GC.....	41
4.7 Cobertura dos genes.....	44
4.8 Variantes não descritas em bancos de dados nos genes de interesse.....	49
5 Discussão.....	51
6 Conclusão.....	55
7 Referências.....	56
Anexos	

Lista de Abreviaturas

ACMG – *American College of Medical Genetics*

ACTH – *Adrenocorticotropic hormone*, ou hormônio adrenocorticotrófico

AVP – Arginina vasopressina

BAM – Formato em compressão binária do arquivo Sequence Alignment/Map

CNV – Copy Number Variation

DNA – Ácido desoxiribonucleico

Dpc – Dias pós coito

FMUSP – Faculdade de Medicina da Universidade de São Paulo

FSH – *Follicle-stimulating hormone*, ou hormônio folículo-estimulante

GATK – Genome Analysis Toolkit

GH – *Growth Hormone*, ou Hormônio do Crescimento

LH – *Luteinizing Hormone*, ou Hormônio Luteinizante

NBDC – *National Bioscience Database Center*

NGS – Next Generation Sequencing

PCR – Polymerase Chain Reaction

SELA – Sequenciamento Em Larga Escala

TSH – *Thyroid-stimulating hormone*, ou Hôrmônio estimulante da tireóide

UCSC – *University of California Santa Cruz*

VCF – *Variant Calling Format*

VUS – *Variant of Uncertain Significance*, ou Variante de Significado Incerto

∉ – Não pertence

≠ – Diferente

= – Igual

Lista de Tabelas

Lista de Tabelas

Tabela 1 – Genes analisados no estudo.....	14
Tabela 2 – Classificação ACMG dos genes encontrados por NGS.....	15
Tabela 3 – Qualidade do DNA das amostras selecionadas.....	19
Tabela 4 – Principais diferenças dos kits utilizados.....	22
Tabela 5 – Quantidade de <i>reads</i> do grupo HapMap.....	27
Tabela 6 – Quantidade de <i>reads</i> do paciente com hipopituitarismo e sua mãe.....	27
Tabela 7 – Quantidade de <i>reads</i> do grupo de amostras aleatórias.....	28
Tabela 8 – Média da nota Phred.....	30
Tabela 9 – <i>Reads</i> mapeados e duplicatas do grupo de amostras aleatórias.....	31
Tabela 10 – Quantidade de genes 100% cobertos do HapMap.....	45
Tabela 11 – Quantidade de genes 100% cobertos do hipopituitarismo.....	45
Tabela 12 – Quantidade de genes 100% cobertos das aleatórias.....	45
Tabela 13 – Cobertura na extensão do gene <i>HES5</i>	46
Tabela 14 – Cobertura na extensão do gene <i>SOX3</i>	48

Lista de Figuras

Figura 1 – Linha do tempo da descoberta de genes relacionados ao hipopituitarismo.....	4
Figura 2 – Etapas envolvidas no sequenciamento.....	6
Figura 3 – Fluxograma do sequenciamento.....	7
Figura 4 – Etapas da análise de bioinformática.....	9
Figura 5 – Gráficos média de cobertura SOX2 e SOX3.....	11
Figura 6 – Heredograma do paciente com hipopituitarismo sequenciados pelos kits Nimblegen e Sure select	17
Figura 7 – Filtragem de variantes pelo ABraOM.....	25
Figura 8 – Gráficos ilustrativos do FASTQC.....	29
Figura 9 – Gráficos de taxa de duplicação.....	32
Figura 10 – Gráficos de saturação global.....	34
Figura 11 – Gráficos de cobertura global.....	35
Figura 12 – Diagrama de Venn das <i>probes</i>	37
Figura 13 – Gráficos de saturação da região de interesse.....	39
Figura 14 – Gráficos de cobertura da região de interesse.....	40
Figura 15 – Gráficos de cobertura da região de GC das amostras HapMap.....	42
Figura 16 – Gráficos de cobertura da região de GC das amostras aleatórias.....	43
Figura 17 – Gráfico de comparação da quantidade de GC nos genes.....	44
Figura 18 – Gráficos de cobertura do gene <i>HES5</i>	47
Figura 19 – Gráficos de cobertura do gene <i>SOX3</i>	48
Figura 20 – Histograma de comparação da quantidade de GC.....	49
Figura 21 – Filtragem de variantes do ABraOM nos genes de interesse.....	50

Resumo

Benedetti AFF. *Análise de diferentes métodos de sequenciamento de larga escala dos genes envolvidos no hipopituitarismo e embriogênese hipofisária* [Dissertação]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2018.

Mutações nos genes envolvidos na embriogênese hipofisária já foram descritas relacionadas a quadros isolados de deficiência hormonal múltipla e/ou associado a fenótipos extra-pituitários. As mutações encontradas em humanos foram descritas em genes envolvidos na embriogênese hipofisária e cujos fenótipos foram gerados em animais a partir de nocaute, servindo de ponto de partida para sua busca em pacientes com fenótipo similar. Essa estratégia é conhecida como busca por gene candidato, e é feita pela técnica de sequenciamento tradicional Sanger. Na última década, com o avanço de novas tecnologias de sequenciamento, diversos genes foram associados ao hipopituitarismo, principalmente utilizando-se a metodologia de exoma. Contudo, ainda há uma grande parcela dessa população sem diagnóstico molecular, como evidenciado em um levantamento na literatura por Fang e colaboradores e cuja tendência foi observada no ambulatório de Endocrinologia do Desenvolvimento do Hospital das Clínicas, onde apenas 14% dos pacientes tiveram o seu diagnóstico molecular determinado. Com isso, as tecnologias de sequenciamento de última geração, passaram a ser uma ferramenta promissora para determinação molecular dos fenótipos dos pacientes. Logo, alguns pacientes em seguimento no ambulatório de endocrinologia do desenvolvimento tiveram o exoma sequenciado, e uma análise das métricas do sequenciamento evidenciou regiões de cobertura muito baixa, o que não permitiu a conclusão sobre a presença ou ausência de variantes nessas regiões. Entre essas regiões estão os genes SOX2 e SOX3, os quais possuem variantes conhecidas causadoras do fenótipo. Esse trabalho tem como objetivo analisar a cobertura dos genes envolvidos na embriogênese hipofisária assim como os relacionados ao hipopituitarismo congênito em quatro diferentes kits de preparação de biblioteca para exoma, a fim de identificar a melhor metodologia para um diagnóstico molecular dos pacientes além determinar variantes específicas da população brasileira na região de interesse através de busca no site ABraOM. Foram analisados 76 genes em um total de 119 amostras separadas em três grupos, sendo o primeiro grupo de amostras HapMap, o segundo de um paciente com hipopituitarismo e sua mãe e o terceiro de amostras brasileiras aleatórias. Os kits utilizados foram NimbleGen (Roche), Nextera (Illumina), SureSelect e SureSelect+UTR (Agilent). Para isso, foram utilizados diversos programas de bioinformática, tendo entre eles o FASTQC, BWA, GATK, Annovar, Qualimap e bedTools. Análises da qualidade do sequenciamento, assim como a taxa de mapeamento e duplicação mostraram que as amostras utilizadas apresentavam qualidades adequadas e similares entre si para a análise. De acordo com os resultados obtidos em relação a cobertura, o kit da NimbleGen apresenta uma queda em sua cobertura dos genes de interesse em relação a sua capacidade de cobertura do exoma global, algo que pode ser devido à alta taxa de GC na região de interesse, uma vez que a capacidade do kit nessas regiões é deficiente em

relação aos demais. Os genes com piores coberturas em todas as quatro tecnologias foram os genes HES5, que apesar de fazer parte da embriologia hipofisária, não possui variante relacionada ao fenótipo em humanos, e o SOX3 que, apesar de ter muita baixa cobertura na NimbleGen, é bem coberto na SureSelect. Isso corrobora com a análise de capacidade de cobertura em regiões com alta taxa de GC. Somado a isso observou-se que a população brasileira tem 885 variantes únicas e exclusivas. Concluimos, portanto, que o kit SureSelect, da Agilent, tem o melhor desempenho na região de interesse, assim como no exoma global, sendo o indicado para estudos em coortes de hipopituitarismo e a população brasileira possui variantes únicas inerentes a ela.

Summary

Benedetti AFF. *Analysis of different methods of high-throughput sequencing of genes involved in hypopituitarism and pituitary embryogenesis* [dissertation]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2018.

Mutations in the genes involved in pituitary embryogenesis have been described related to isolated cases of multiple hormonal deficiency and/or associated with extra-pituitary phenotypes. Mutations found in humans were described in genes involved in pituitary embryogenesis by generating phenotypes knockout animals, serving as the starting point for their search in patients with similar phenotype. This strategy is known as gene candidate search and is performed by the traditional Sanger sequencing technique. In the last decade, with the advancement of new sequencing technologies, several genes have been associated with hypopituitarism, mainly using the exome methodology. However, there is still a large portion of this population without molecular diagnosis, as evidenced by a survey in the literature by Fang *et al.*, This trend was also observed in the outpatient clinic of Developmental Endocrinology of Hospital das Clínicas, where only 14% of the patients had their molecular diagnosis. With this, high throughput sequencing technologies have become a promising tool for the molecular determination of patients' phenotypes. Therefore, we sequenced the exome of some of our patients, and an analysis of the sequencing quality showed very low coverage regions, which harms the researcher's ability to reach a conclusion regarding presence or lack of variants in these regions. Among these regions are the SOX2 and SOX3 genes, which have many variants that are known to cause the phenotype. This work aims to analyze the coverage of the genes involved in pituitary embryogenesis as well as those related to congenital hypopituitarism in four different exome library preparation kits in order to identify the best methodology for a molecular diagnosis of the patients and to determine specific variants of the Brazilian population in the region of interest by searching the ABraOM website. A total of 76 genes were analyzed in a total of 119 samples in three groups: the first group of HapMap samples, the second of a patient with hypopituitarism and his mother, and the third group of random Brazilian samples. The kits used were NimbleGen (Roche), Nextera (Illumina), SureSelect and SureSelect + UTR (Agilent). For this, several bioinformatics programs were used, among them FASTQC, BWA, GATK, Annovar, Qualimap and bedTools. Sequencing quality analysis, as well as the mapping and duplication rate, showed that the samples used presented adequate and similar qualities for the comparison. According to the results obtained in relation to the coverage, the NimbleGen kit shows a drop in its coverage of the genes of interest in relation to its capacity to cover the global exome, something that may be due to the high GC rate in the region of interest, once the capacity of the kit in these regions is not as good as the others. The genes with the worst coverage in all four technologies were the *HES5* gene, which despite being part of the pituitary embryology, have no phenotype-related variant in humans, and SOX3 which, despite having very low coverage in NimbleGen, is well covered on SureSelect. This corroborates the analysis of coverage capacity in regions with a high GC rate. In addition to this it was observed that the Brazilian population has 885 unique and exclusive variants. Therefore, we conclude that the Agilent's SureSelect kit has the best performance in the region of interest, as well as in the global exome, being recommended for studies in

hypopituitarism cohorts, and that the Brazilian population has unique variants inherent to it.

1. Introdução

1. INTRODUÇÃO

1.1. Embriogênese hipofisária

A hipófise localiza-se na sela túrcica, logo abaixo do hipotálamo, no qual se encontra fisicamente ligada pela haste hipofisária. É uma glândula formada por três lobos: anterior, intermediário (ambos formando a adenohipófise), e o posterior (ou neurohipófise)¹⁻³. A adenohipófise, originada do ectoderma oral, é composta por somatotrofos, tireotrofos, gonadotrofos, corticotrofos e lactotrofos, que produzem os hormônios GH, TSH, FSH e LH, ACTH, e prolactina, respectivamente, além dos melanotrofos, localizados no lobo intermediário, produtores de α MSH¹. A neurohipófise, originada no ectoderma neural, secreta arginina-vasopressina (AVP) e ocitocina a partir dos estímulos hipotalâmicos^{1,2}.

O desenvolvimento da glândula pituitária dos camundongos é um bom modelo animal para entender o desenvolvimento pituitário em humanos³. Em camundongos, os estágios de desenvolvimento da hipófise se iniciam no primeiro dia pós-coito (dpc) e os tempos marcantes podem ser resumidos em: 9,0dpc, no qual o placóide invagina em direção ao diencéfalo, formando assim a bolsa de Rathke. No 10 dpc, a região do diencéfalo ventral diretamente acima a Bolsa de Rathke começa a evaginar em sua direção e forma o infundíbulo, o qual dará origem ao hipotálamo, à haste hipofisária e à neurohipófise^{1,4,5}. Entre 10,5 e 12,5dpc, a bolsa se fecha e se separa do ectoderma oral adjacente^{4,5}. A partir de 12,5 e 15,5dpc, se inicia a proliferação de células progenitoras que formarão o lobo anterior da hipófise. As células na região dorsal da bolsa

formarão o lobo intermediário, e o lúmen da bolsa será a fenda hipofisária, que separa os lobos anterior e intermediário^{4,5}.

O desenvolvimento adequado da glândula depende da expressão temporal e espacial de fatores de transcrição e moléculas sinalizadoras⁶. Expressão ou função inadequada desses fatores causa um desenvolvimento deficiente da glândula, que leva a diminuição ou falta de secreção dos hormônios pituitários contribuindo para o diagnóstico clínico de hipopituitarismo congênito⁷.

Durante a invaginação do teto da cavidade oral seguido pela formação da bolsa de Rathke, são expressas moléculas sinalizadoras no diencéfalo conhecidas como fatores de crescimento do fibroblasto (*FGF8*, *FGF10* e *FGF18*), do *sonic hedgehog* (*SHH*), das proteínas morfogênicas do osso (*BMP2* e *BMP4*) e das WNTs (*WNT4*, *WNT5A*), tendo sido descritas mutações em alguns desses genes (*FGF8*, *SHH*) associadas à deficiência pituitária múltipla⁷.

1.2. Hipopituitarismo congênito

O hipopituitarismo é uma condição que afeta 1 em cada 8000 indivíduos no mundo todo, e se caracteriza pela deficiência de um ou mais hormônios da adenohipófise. Considerando que a deficiência de GH está sempre presente, isso afeta o desenvolvimento de diversas partes do corpo⁸. Enquanto os primeiros sinais costumam ser relacionados à baixa estatura, o paciente afetado também pode apresentar ganho de peso e fadiga, ausência ou atraso da puberdade e infertilidade, dependendo de quais hormônios não estão sendo produzidos pela glândula⁸. Quando o quadro é congênito, pode-se apresentar

variações, indo de deficiência de GH isolada ou associada a deficiências combinadas e má formações como holoprosencefalia e displasia septo-óptica, sendo então uma doença de largo espectro que pode ser sindrômica (quando relacionada a fenótipos mais graves) ou não sindrômica (relacionado apenas a deficiências hormonais)⁹.

Mutações nos genes *HESX1*, *SOX2*, *SOX3*, *LHX3*, *LHX4*, *OTX2*, e *GLI2*, já foram descritas em associação com fenótipos extra-pituitários, como anormalidades craniofaciais, displasia septo-óptica e holoprosencefalia⁷. Já mutações em fatores de transcrição expressos em estágios finais de diferenciação das células hipofisárias, como *PROP1* e *POU1F1*, foram descritos em humanos não estando associadas a outras más formações. Essas mutações foram descritas a partir de fenótipos produzidos em animais que serviram de ponto de partida para sua busca em pacientes com deficiências hormonais devido a semelhanças em fenótipos das duas espécies. Essa estratégia é conhecida como busca do diagnóstico molecular a partir de gene candidato, que é feita pela técnica de sequenciamento tradicional denominada de Sanger⁹. Nos últimos 5 anos, com o aparecimento de novas tecnologias de sequenciamento, diversos genes foram associados ao hipopituitarismo, como *IGSF1*, *ARNT2*, *TGIF1*, *CHD7*, *PNPLA6*, *DMXL2*, *KAL1*, *ZSWIM6*, *IFT172*, *HNRNPU*, *GPR161*, *CDON*, *NFKB2*, *KCNQ1*, *FOXA2*, *SLC15A4* e *SCL20A1* que antes não seriam supostos associados ou causadores dessa doença⁹⁻¹¹ (Figura 1).

Figura 1. Linha do tempo ilustrando a descoberta dos genes pelo método de Sanger no total de 15 genes em 30 anos e a associação de 9 novos genes num período de 6 anos pelo sequenciamento de nova geração.

1.3. Sequenciamento por metodologia de Sanger

A tecnologia de Sanger para sequenciamento de DNA, utiliza fluorescência em oligonucleotídeo terminador (didexonucleotídeo), eletroforese capilar e detecção por sinalização de laser para a análise de sequências de 500 a 1000 pares de bases, sendo então utilizada na rotina de sequenciamento de pequenos fragmentos de DNA^{12,13}. Apesar de ser uma técnica que demanda grande quantidade de tempo e reagentes, é o “gold standard” para sequenciamentos, sendo ainda muito utilizada em laboratórios de diagnóstico clínico e pesquisa básica, assim como para validar resultados encontrados por novas técnicas de sequenciamento de nova geração.

1.4. Sequenciamento de nova geração

O sequenciamento de nova geração (NGS, na sigla em inglês), também conhecido como sequenciamento de larga escala, se torna cada vez mais rotineiro em laboratórios clínicos e de pesquisa básica^{14,15}. O avanço tecnológico destes tipos de plataformas permitiu a produção de maior

quantidade de dados no sequenciamento de um indivíduo, possibilitando a identificação de novas ou raras variantes, assim como a diminuição do custo dessa tecnologia ao longo dos anos¹⁶. Além da abordagem de permitir a identificação de variantes raras para doenças com herança mendeliana, também investiga bases genéticas de doenças que não são exclusivamente monogênicas, apontando variantes e polimorfismos relacionados ao risco da manifestação da doença^{17,18}. Dentre as tecnologias de sequenciamento de última geração, englobam-se três metodologias referentes a diferentes partes do genoma sequenciado, sendo elas o genoma, o exoma e o painel de genes conhecidos. Todas as três utilizam o mesmo princípio e sua principal diferença é a porção do DNA sequenciado, sendo, respectivamente, o genoma completo, a parte exônica e codificadora de proteínas, que corresponde a 2-3% do genoma, e genes de interesse do pesquisador ¹⁴. São metodologias adequadas para a descoberta de variantes e mudanças estruturais, como CNVs (copy number variation), inserções e deleções ^{19,20} em um grande número de genes de uma grande casuística e vem sendo utilizadas principalmente em estudos de características genéticas complexas permitindo a descoberta sistemática, previamente impossível, de variantes raras em milhares de amostras e a associação dessas variantes com o fenótipo de interesse^{14,21,22}. Dentre as tecnologias de sequenciamento apresentadas, o exoma global, o qual se propõe a analisar a região exônica, incluindo regiões de UTRs e microRNAs, torna-se uma estratégia eficiente na busca de novas variantes alélicas com impacto funcional²³⁻²⁵, visto que para doenças mendelianas 85% das variantes alélicas deletérias se encontram inseridas nas regiões codificadoras ^{26,27}.

As principais etapas do exoma são: a preparação da biblioteca, ou seja, preparação da amostra para uma forma que seja compatível com o sistema, seguido pelo sequenciamento de elevada capacidade e, por fim, a análise de bioinformática dos dados gerados ^{28,29} (Figura 2).

Figura 2. Etapas envolvidas no processo de sequenciamento (adaptado de Rehm, 2013³⁰), representando o sequenciamento por genoma (em setas pretas), e sequenciamento por exoma e painel de genes (em setas tracejadas).

Para o sequenciamento do exoma global, é necessária a preparação de uma amostra com pequenos fragmentos de DNA (de 100 a 500 pares de bases) flanqueados por adaptadores específicos³⁰. Esse processo envolve a fragmentação e/ou dimensionamento das sequências de interesse para o comprimento desejável, que pode ser feita por meios mecânicos ou

enzimáticos. Posteriormente, é feita a junção de adaptadores de oligonucleotídeos ao final dos fragmentos, além da junção dos *barcodes*, que são sequências de três ou mais pares de bases que servirão como marcadores específicos para os fragmentos de DNA, permitindo que diversas amostras sejam sequenciadas juntas. Esse processo se segue a uma seleção por *beads* magnéticas para que se tenha uma distribuição adequada dos tamanhos e posteriormente, para uma amplificação por PCR dos fragmentos e quantificação do produto final para verificar a adequação da amostra para o sequenciamento^{29,30}.

Figura 3. Fluxograma dos passos de execução da preparação da biblioteca, seguido pelo sequenciamento e o tratamento dos dados pela bioinformática. (Adaptado de Bamshad *et al.*, 2011³¹)

Diversas empresas disponibilizam no mercado kits para a construção de bibliotecas a partir do DNA cujas diferenças são referentes à seleção da região alvo, método de fragmentação das moléculas, tamanho e densidade das *bait*s, que é a molécula utilizada para a captura, seguida pelo processo de hibridação^{29,32}. As diferenças nos reagentes utilizados e a forma de detecção das bases sequenciadas podem ser responsáveis por vieses no

sequenciamento, resultando em diferenças na quantidade total de bases/fragmentos gerados, tamanho dos fragmentos lidos, tempo total do procedimento e, por fim, a qualidade e a acurácia dos dados finais^{29,30}.

Após essa etapa inicial, a amostra será sequenciada na plataforma de escolha, as quais se diferem entre o método de sequenciamento, podendo ser por ligação ou síntese, e no método de detecção de bases sequenciadas, que pode ser por fluorescência, luminescência ou diferença de íons H⁺³³. A leitura do fragmento pode ser realizada em apenas uma das pontas dos fragmentos (chamado *single-end*) ou de ambas as pontas (*paired-end*)³⁰. Devido a geração de grande quantidade de dados das sequências, é necessária uma análise de bioinformática para tratar os dados gerados pelo sequenciador, que é feito através da chamada de bases, alinhamento dos *reads* ao genoma de referência e, por fim, da chamada e anotações das variantes^{30,34–36}. Para uma chamada de variantes eficiente, é necessário que o sequenciamento possua um mínimo de cobertura, tanto das extensões dos exons (chamada de cobertura horizontal), quanto das bases (calculada pela quantidade de diferentes *reads* que cobrem a mesma base, chamada de profundidade ou cobertura vertical), sendo considerados mínimos de 99% e 20x respectivamente^{18,36,37}. (Figura 4)

Figura 4. Etapas da análise de bioinformática, adaptada de Altmann³⁴.

Com a grande quantidade de dados genéticos gerados, propiciados pelo avanço destas tecnologias, diversos genes que antes não seriam estudados em pacientes com hipopituitarismo congênito foram encontrados em seu material genético, principalmente utilizando-se da metodologia de exoma. Contudo, ainda há uma grande parcela dessa população que não possuem diagnóstico molecular, como evidenciado em um levantamento na literatura por De Rienzo e colaboradores³⁸ e Fang e colaboradores⁹, os quais mostram que aproximadamente 15,8% dos pacientes tiveram o diagnóstico molecular determinado em genes envolvidos na embriogênese hipofisária. Essa tendência foi observada no ambulatório de Endocrinologia do Desenvolvimento do Hospital das Clínicas, onde apenas 14% dos pacientes tiveram o seu diagnóstico molecular determinado. Diante desse cenário, optou-se por aplicar

o exoma no DNA de alguns dos pacientes em seguimento na Unidade de Endocrinologia do Desenvolvimento, em parceria com a Prof^a Sally Camper, utilizando o kit de preparação NimbleGen SeqCap EZ v3 produzido pela Roche (Roche Inc., Madison, WI, USA), sendo realizados no *Sequencing Core* da Universidade de Michigan, Ann Arbor, MI, USA.

Entretanto, uma análise das métricas desses sequenciamentos, que implica na verificação da qualidade de cobertura, apontou regiões deficientes da mesma, tendo como exemplo os genes *SOX2* e *SOX3* (Figura 5), ambos com diversas variantes relacionadas ao hipopituitarismo na literatura^{39–44}. Essa constatação pode acarretar na impossibilidade de se chegar a uma conclusão sobre a presença ou não de variantes nessas regiões, o que levou ao processo laborioso de se avaliar por Sanger os genes *SOX2* e *SOX3* desses pacientes enviados para o exoma. Diante deste panorama duas hipóteses para explicar a baixa cobertura surgiram: i-) A baixa cobertura está atrelada ao kit de captura utilizado? Ou ii-) Está associada às regiões, como por exemplo as ricas em GC, que sabidamente dificultam sua captura independente do kit utilizado?⁴⁵ Para responder à essas perguntas foi proposto esse projeto.

Figura 5. Gráficos ilustrando a cobertura média dos genes SOX2 e SOX3 nos pacientes submetidos ao sequenciamento pelo kit NimbleGen. A linha vermelha indica a faixa de 20x de cobertura. Conclui-se que 28,5% da extensão do SOX2 assim como mais de 56,6% do SOX3 estiveram abaixo da cobertura mínima, representado pela linha vermelha. Os gráficos foram feitos utilizando a função `plot.baseCoverage`, disponível pelo link (<https://github.com/anna-benedetti9/>).

2. Objetivos

2. Objetivos

2.1. Primário

Analisar a cobertura dos genes envolvidos na embriogênese hipofisária e relacionados ao hipopituitarismo congênito em humanos em quatro diferentes kits de preparação de biblioteca, a fim de identificar o melhor dentre eles para ser utilizado como ferramenta de diagnóstico molecular dos pacientes.

2.2. Secundários

- Fazer uma busca na literatura para identificar os genes associados à embriogênese hipofisária em modelos animais e/ou associado ao quadro de hipopituitarismo congênito em humanos, a fim de usá-los como genes alvo da métrica.
- Analisar o número de variantes alélicas exclusivas da população brasileira utilizando o banco de dados ABRAOM comparado com outros bancos internacionais.

2.3 Considerações éticas

Esse estudo foi submetido à análise do comitê de ética médica e os NGS realizados tiveram diferentes protocolos previamente assinados para a realização do mesmo. CEP/CONEP colaboração com a Universidade de Michigan: CAAE 06425812.4.0000.0068.

3. Metodologia

3. Metodologia

3.1. Pesquisa bibliográfica de genes a serem incluídos no estudo

A pesquisa feita na literatura teve como escopo a busca de genes relacionados ao hipopituitarismo e publicados até junho de 2018 (Tabela 1). As variantes foram classificadas de acordo com os critérios de patogenicidade da American College of Medical Genetics (ACMG)³⁰, utilizando a ferramenta Genetic Variant Interpretation Tool disponível através do site da Universidade de Maryland (http://www.medschool.umaryland.edu/Genetic_Variant_Interpretation_Tool1.html). Os genes analisados foram incluídas no estudo independentemente do nível de evidência (Tabela 2). Além desses genes foram incluídos genes que possuem papel na embriogênese hipofisária porém nunca foram associados ao fenótipo de hipopituitarismo. Selecionamos um total de 76 genes, sendo 44 associados ao hipopituitarismo em humanos e 32 relacionados à embriogênese hipofisária, mas que não possuem mutações relacionadas ao fenótipo humano^{3,9}.

Tabela 1. Genes relacionados ao hipopituitarismo e a embriogênese hipofisária.

Relacionados ao Hipopituitarismo		Relacionados à Embriogênese Hipofisária	
ARNT2	KCNQ1	AES	SIX3
CDH2	LEPR	BMP2	SIX6
CDON	LHX3	BMP4	SOX4
DMXL2	LHX4	BMP7	TBX2
FGF8	NFKB2	BMPR1A	TBX3
FGFR1	OTX2	CHD7	TCF4
FOXA2	PAX6	CTNNB1	TLE1
GH1	PITX2	FGF10	WNT4
GHR	PNPLA6	FGF18	ZSWIM6
GHRH	POU1F1	GATA2	
GHRHR	PROKR2	GATA3	
GHSR	PROP1	GLI3	
GLI2	RNPC3	HES1	
GPR161	SHH	HES5	
HDAC6	SLC15A4	HNRNPU	
HESX1	SLC20A1	ISL1	
HHIP	SOX2	LHX2	
IFT172	SOX3	NOTCH2	
IGSF1	TCF7L1	POLR3A	
IRF6	TGIF1	RAX	
JAK1	WDR11	RBM28	
KAL1	WNT5A	RBPJ	

Tabela 2. Classificação ACMG dos genes relacionados ao hipopituitarismo encontrados por NGS.

GENE	VARIANTE	CLASSIFICAÇÃO ACMG	REFERÊNCIA
ARNT2	c.1373_1374dupTC/p.Y459Rfs*52	Provavelmente patogênica	Webb <i>et al</i> , 2013 ⁴⁶
ZSWIM6	p.R1163W	Provavelmente patogênica	Smith <i>et al</i> , 2014 ⁴⁷
PNPLA6	Várias	VUS	Synofzik <i>et al</i> , 2014 ⁴⁸ , Hufnagel <i>et al</i> , 2015 ⁴⁹
HNRNPU	c.1615-1G>A	VUS	Zhu <i>et al</i> , 2015 ⁵⁰
GPR161	c.56T>A/p.Leu19Gln	VUS	Karaca <i>et al</i> , 2015 ⁵¹
CDON	c.2764T>C/Glu922Ter	VUS	Bashamboo <i>et al</i> , 2016 ⁵²
CHD7	c.2194C>G/p.P732A	VUS	Gregory <i>et al</i> , 2013 ⁵³
IFT172	c.5179T>C/p.Cys1727Arg; c.337-2A>C	VUS	Lucas-Herald <i>et al</i> , 2015 ⁵⁴
DMXL2	p.1942_1946del	VUS	Tata <i>et al</i> , 2014 ⁵⁵
KAL1	p.His568Gln	VUS	Takagi <i>et al</i> , 2014 ⁵⁶
KCNQ1	c.347G>T/p.(Arg116Leu); c.1106C>T/p.(Pro369Leu)	Provavelmente patogênica	Tommiska <i>et al</i> , 2017 ⁵⁷
IRF6	c.697C>T/p.Arg233Cys	VUS	Starink <i>et al</i> , 2017 ⁵⁸
NFKB2	c.2596A>C/p.Ser866Arg	VUS	Lal <i>et al</i> , 2017 ⁵⁹
FOXA2	c.505T>C/p.S169P	Provavelmente patogênica	Giri <i>et al</i> , 2017 ⁶⁰
JAK1 LEPR	8Mb Deletion of Ch 1p31.1 - 1p31.3	VUS VUS	Thakur <i>et al</i> , 2017 ⁶¹
SLC15A4	p.P456L; p.L84F	Provavelmente patogênica	Simm <i>et al</i> , 2017 ⁶²
SLC20A1	p.L89S; p.L521F	Provavelmente patogênica	
MAGEL2*	c.3019 C > T/p.Gln1007Ter	VUS	Hidalgo-Santos <i>et al</i> , 2018 ⁶³

* Gene não presente na análise pois foi publicado após a última pesquisa feita em Junho/2018

VUS: Variante de significado incerto, do inglês *Variant of Uncertain Significance*

3.2. Amostras

Foram utilizados 3 grupos distintos de amostras para a comparação das tecnologias, como segue:

3.2.1. Amostras providas do projeto HapMap

O projeto HapMap foi criado para mapear haplótipos do genoma humano, utilizando-se de amostras de diversas populações. Estas foram

posteriormente disponibilizadas para outros pesquisadores, e foram bastante utilizadas em artigos de comparação de técnicas por possuírem variantes bem caracterizadas^{32,64,65}. O Laboratório denominado *Genotyping Development*, do *RIKEN Center for Integrative Medical Sciences*, em Yokohama no Japão sequenciou duas amostras NA18943 e NA18948, sendo ambas masculinas, de origem japonesa e sem anormalidades de cariótipo. Essas amostras foram sequenciadas em três diferentes tecnologias como segue: Nextera, NimbleGen e SureSelect v5. Foram depositadas no National Bioscience Database Center (NBDC) Human Database (<http://humandbs.biosciencedbc.jp/>), e o seu download foi feito através do link (<http://trace.ddbj.nig.ac.jp/DRAsearch/submission?acc=DRA003736>).

3.2.2. Amostras do paciente com hipopituitarismo e sua mãe

Um paciente brasileiro foi sequenciado no Sequencing Core da Universidade de Michigan utilizando o kit da NimbleGen e no Laboratório de Sequenciamento de Larga Escala (SELA) a *facility* de multiusuários da Faculdade de Medicina da USP (FMUSP) (projeto FAPESP nº 14/50137-5), com o kit SureSelect para comparação entre as duas tecnologias. Esse paciente é filho de pais consanguíneos, primos de segundo grau, e veio ao Hospital das Clínicas com queixa de baixa estatura desde os 7 anos e 2 meses (7a2m) de idade. Na época, ele tinha 89cm (Z de Altura -5,89) e uma idade óssea atrasada de 2a. Após ter sido feito o diagnóstico de deficiência de GH, foi iniciado o tratamento com GH aos 9a2m, com 92,5 cm (Z Alt -6,8) e idade óssea de 2a6m. A ressonância magnética de hipófise evidenciou adenohipófise hipoplásica, com haste ausente e neurohipófise ectópica.

Aos 11 anos, evoluiu com deficiência de TSH e foi introduzido a levotiroxina. Teve início de puberdade espontâneo aos 16a10m, mas foi bloqueada até os 18a9m, quando estava com 139,4cm (Z Alt -4,5). Aos 19a8m teve a puberdade induzida pois não apresentava caracteres sexuais secundários, concluindo-se que o mesmo evoluiu com deficiência das gonadotrofinas. Atingiu altura final de 163cm aos 28a6m, próximo a sua altura alvo de 165,25 (Z Alt -1,42). Sua mãe fenotipicamente normal foi encaminhada para o exoma (Figura 6).

Figura 6. Heredograma do paciente com hipopituitarismo. A seta indica o paciente índice. Foi sequenciado ele e sua mãe.

3.2.3. Amostras aleatórias

Foram sequenciadas no *Sequencing Core* da Universidade de Michigan 55 amostras de pacientes com hipopituitarismo e seus familiares. No SELA foram sequenciadas 628 amostras e esse foi denominado banco interno. Critérios de inclusão: amostras de DNA extraídos de sangue periférico. Critérios de exclusão: i) Amostras de DNA obtidas a partir de mucosa oral, e ii)

Amostras cujo arquivo de mapeamento (BAM) estavam incompletos, não preenchendo os critérios de qualidade. Foram selecionadas randomicamente um total de 109 amostras para a análise final, somando 43 da NimbleGen (Roche Inc.), 20 da Nextera (Illumina Inc.), 26 da SureSelect e 20 da SureSelect com UTR (Agilent Inc).

3.3. Extração de DNA

As amostras de DNA dos pacientes selecionados e familiares foram obtidas a partir de leucócitos de sangue periférico e extraídas pela técnica de *Salting out* de acordo com procedimento padronizado no Laboratório de Hormônios e Genética Molecular da Unidade de Endocrinologia do Desenvolvimento, da FMUSP⁶⁶.

A concentração do DNA extraído foi obtida por leitura em espectrofotômetro no comprimento de onda de 260 nm (1 unidade DO 260 = 50 µg/mL). A relação ideal entre as leituras em 260 e 280 nm para a caracterização da pureza do material deve ser superior a 1,75. As amostras são mantidas congeladas a -20 °C até seu uso. É possível verificar uma consistência na relação A260/A280 nas amostras utilizadas, indicando boa qualidade das amostras selecionadas, na Tabela 3.

Tabela 3. Qualidade do DNA das amostras selecionadas, indicando a relação A260/A280 e sua concentração.

Amostra	Relação		Amostra	Relação	
	A260/A280	[μ g/ml]		A260/A280	[μ g/ml]
1210	1,9	71,1	7709	1,8	355
1211	1,8	135	7846	1,8	266
1212	1,8	83,3	9378	1,8	289
1629	1,8	186	10266	1,9	164
1985	1,8	71,1	10317	1,7	174
2053	1,8	287	10460	1,9	176
2078	2	49	6950	1,8	85,9
2260	2	44,1	6951	1,9	46,6
2261	1,8	132	6952	1,9	118
2262	1,8	103	1493	1,7	1363
10654	1,8	262	1494	1,8	208
10655	1,8	228	3028	2	123
10656	1,8	78,4	3029	1,8	191
10867	1,8	240	334N	1,9	21
10868	1,9	211	335	1,9	29,5
11004	1,7	29,4	336	2	14
11005	1,6	80,9	10652	1,9	30,5
11088	1,8	63,7	10658	2	18
116	2	53,9	3009	1,8	14,5
2298	1,8	242	3010	1,8	22,5
6599	1,8	137	4012	1,8	166
6600	1,7	127	4013	1,8	278
8501	1,8	583	4014	1,8	312
8502	1,8	287	4218	1,8	162
8503	1,7	123	4925	1,8	392
7920	1,8	172	6980	1,8	146
10427	1,8	177	7589	1,8	249
10428	1,8	194	7647	1,8	155
1569	1,8	196	7842	1,7	497
6615	1,8	184	7843	1,7	459
8716	1,8	373	8012	1,8	153
10875	1,9	150	8013	1,7	150
10876	1,9	90,7	8016	1,7	189
10877	1,8	157	8017	1,7	544
273	NA	NA	8609	1,7	84
797	1,8	301	8859	1,7	58
1075	1,9	136,9	8860	1,7	282
1291	1,8	153	8862	1,7	210
1292	1,8	161	8863	1,7	151
1293	1,9	268	10060	1,8	164
1294	1,8	311	10071	1,8	375
2172	1,8	451	2715	1,8	152
3428	1,8	232	3419	1,8	143

3.4. Preparação das bibliotecas

3.4.1. Nextera

O sequenciamento de NGS pela metodologia Nextera foi realizado no SELA (FMUSP – São Paulo, Brasil). As amostras do DNA genômico foram submetidas à fragmentação enzimática conhecida como tagmentação, que se caracteriza pela quebra do DNA em aproximadamente 300 pb com a adição simultânea das sondas e amplificação dos fragmentos. Posteriormente, foi realizada a captura dos fragmentos por *bait*s de DNA e o enriquecimento das regiões codificadoras dos genes utilizando o sistema de reagentes Nextera ocorreu de acordo com os protocolos do fabricante (Illumina Inc.). As amostras que atenderam ao controle de qualidade foram submetidas ao sequenciamento na plataforma HiSeq 2500 (Illumina Inc., San Diego, CA, USA) com *paired-ends* como configuração.

3.4.2. NimbleGen

As amostras do DNA genômico foram submetidas à fragmentação mecânica usando um ultrasonicador da Covaris, no *Sequence Core* da Universidade de Michigan. Em seguida, foram realizadas a captura dos fragmentos por *bait*s de DNA e preparadas as bibliotecas destes fragmentos utilizando-se o sistema automatizado Apollo 324 e kits de reagentes específicos de acordo com os protocolos recomendados pelo fabricante (IntegenX Inc., Pleasanton, CA, USA). Esta etapa é responsável pelo isolamento e purificação dos fragmentos, reparo da ponta, ligação aos adaptadores e seleção dos fragmentos do DNA genômico de acordo com o tamanho desejado. Posteriormente, foi realizada a captura e enriquecimento das regiões codificadoras utilizando o sistema e

reagentes NimbleGen SeqCap EZ v3 de acordo com os protocolos do fabricante (Roche Inc., Madison, WI, USA). As amostras que atenderam ao controle de qualidade foram submetidas ao sequenciamento na plataforma HiSeq 2000 (Illumina Inc., San Diego, CA, USA) na configuração de *paired-end reads*, gerando leituras de 100 nucleotídeos em cada extremidade dos fragmentos.

3.4.3. SureSelect/SureSelect + UTR

O sequenciamento do exoma pela metodologia SureSelect foi realizado no SELA. As amostras do DNA genômico foram submetidas à fragmentação mecânica usando um ultrasonicador da Covaris. Em seguida, os fragmentos foram capturados por *baits* de RNA e preparadas as bibliotecas destes fragmentos utilizando-se kits de reagentes específicos de acordo com os protocolos recomendados pelo fabricante (Agilent Inc. Santa Clara, CA, USA). Esta etapa foi responsável pelo isolamento e purificação dos fragmentos, reparo da ponta, ligação aos adaptadores e seleção dos fragmentos do DNA genômico de acordo com o tamanho desejado. Posteriormente, foi realizada a captura e enriquecimento das regiões codificadoras utilizando o sistema de reagentes SureSelect v5 ou SureSelect + UTR v5 de acordo com os protocolos do fabricante (Agilent Inc. Santa Clara, CA, USA). As amostras que atenderam ao controle de qualidade foram submetidas ao sequenciamento na plataforma HiSeq 2500 (Illumina Inc., San Diego, CA, USA), com *paired-end reads*.

Para facilitar a visualização das características dos diferentes kits, disponibilizamos na tabela 4, uma comparação dos diferentes mecanismos de

fragmentação, tipos de sondas e adaptadores, além do tamanho da região alvo.

Tabela 4. Principais diferenças entre os kits utilizados.

TECNOLOGIA	REGIÃO ALVO	TIPOS DE SONDAS	ADIÇÃO DO ADAPTADORES	TIPO DE FRAGMENTAÇÃO
NEXTERA	45 Mb	DNA	Transposase	Enzimático
NIMBLEGEN	64 Mb	DNA	Ligamento	Mecânico
SURESELECT	50 Mb	RNA	Ligamento	Mecânico
SURESELECT+UTR	75 Mb	RNA	Ligamento	Mecânico

3.5. Análise bioinformática

Como representado na figura 4, todas as amostras foram armazenadas no formato FASTQ, que possui informações de todas as bases sequenciadas atreladas com a qualidade de seu sequenciamento. Para averiguar a qualidade foram utilizados diversos programas de bioinformática. O primeiro deles é o programa FASTQC⁶⁷, que é responsável pela análise de controle de qualidade das sequências através da nota Phred, que aponta a probabilidade da base na sequência ser realmente uma base verdadeira, sendo considerada uma boa nota aquela acima de 30^{36,68}. Em seus resultados representados por diversos gráficos, é possível analisar a qualidade das bases, como também a proporção de nucleotídeos gerados por posição, além das sequências hiper-representadas nos fragmentos, detecção de adaptadores nos fragmentos, distribuição de GC e uma predição da porcentagem de fragmentos duplicados. Com esta primeira abordagem, é possível verificar a presença de viés na amostra inicial ou um viés técnico durante o sequenciamento⁶⁷.

Após a checagem da qualidade do sequenciamento, as amostras foram mapeadas utilizando o algoritmo BWA⁶⁹ contra o genoma humano de referência hg19 depositado no UCSC (University of California Santa Cruz)⁷⁰. Este algoritmo possui capacidade de mapear várias pequenas sequências em um curto período, no entanto, pode criar viés de mapeamento em regiões de alta variabilidade. O arquivo BAM é formado após o alinhamento, juntamente com um arquivo indexado para facilitar a pesquisa nele (BAI), que atualmente é o arquivo de saída padrão para este e outros programas de alinhamento⁶⁹.

Os arquivos de mapeamento (BAM) são utilizados para a análise de cobertura e profundidade dos genes, por programas como o bedTools⁷¹ e o Qualimap^{72,73}. O primeiro, baseado no algoritmo do UCSC Genome Browser, possui diversas ferramentas com utilidades para análises de resultados de sequenciamento, incluindo o coverageBed, ferramenta capaz de retornar número de reads por base da região de interesse, em um arquivo conciso de extensão BED⁷⁰. Já o segundo programa, Qualimap, analisa a contagem de reads das regiões mapeadas dentro e fora (*off-target*) da região de interesse de forma geral e mostra seus resultados através de gráficos, além de um sumário geral de todos os dados da amostra^{72,73}.

Após o mapeamento, o algoritmo SAMTools⁷⁴ foi utilizado para a remoção de duplicatas de PCR, uma vez que estes fragmentos podem introduzir artefatos nas análises posteriores, assim como superestimar a cobertura das regiões³⁴.

O software GATK (Genome Analysis Toolkit)⁷⁵⁻⁷⁷ foi utilizado para identificação de variantes de uma única base e inserções ou deleções no DNA em relação ao genoma referência, dando origem ao arquivo tipo VCF (variant

calling format). Nesta etapa foram introduzidos três passos, o primeiro contempla o realinhamento local para ajustar o mapeamento dos *reads*, o segundo de recalibragem de bases e terceiro de recalibragem de variantes chamadas, baseando-se nos bancos de dados HapMap3, 1000 Genome Project⁷⁸, OMIM⁷⁹ e dbSNP⁸⁰ para identificar o perfil dos parâmetros medidos nas variantes conhecidas.

A anotação funcional das variantes foi realizada com o ANNOVAR⁸¹ que utiliza informações atualizadas para anotar essas variantes encontradas pelo GATK, caracterizando-as sobre seu impacto e tipo (nonsynonymous, synonymous, missense, frameshift, splicing-site), assim como possíveis mudanças de aminoácido. É capaz de indicar se essas alterações são deletérias ou possuem um alto impacto, que pode estar relacionado à sua localização e a sua frequência na população em relação aos projetos 1000GenomeProject, EXAC^{82,83} e ESP6500⁸⁴, assim como a conservação de bases entre as espécies^{76,81}.

3.6. ABraOM

ABraOM é um banco de dados que contém informações de variantes genômicas brasileiras. Foi criado a partir de um estudo no qual se fez o exoma de 609 indivíduos idosos. Naslavsky e colaboradores⁸⁵ encontraram 1.282.008 variantes nesses indivíduos e fizeram uma filtragem de qualidade e presença em bancos de dados internacionais. Identificaram nessa comparação 207.621 variantes exclusivas da população brasileira, ou seja, que não haviam sido previamente descritas. Dessas variantes, 46.687 foram encontradas em regiões codificadoras ou de splicing⁸⁵ (Figura 7). Esse é o primeiro banco de

dados de uma população brasileira, que é uma população com alta miscigenação de acordo com Pena e colaboradores⁸⁶.

Figura 7. Filtragem de variantes pelo ABraOM, adaptado de Naslavsky *et al*⁸⁵.

4. Resultados

4. Resultados

Os resultados apresentados a seguir abordam inicialmente os aspectos gerais do sequenciamento nos três grupos de amostras (Grupo1 - HapMap, Grupo2 - paciente com hipopituitarismo e mãe e Grupo3 - amostras aleatórias).

No total, foram analisadas 119 amostras dos 3 grupos, sendo 47 preparadas pelo kit da NimbleGen, 22 pela Nextera, 30 pela SureSelect e 20 SureSelect+UTR. As quantidades médias de *reads* por amostra, separada em 3 grupos, para cada uma das tecnologias podem ser vistas nas Tabelas 5 (Grupo 1), 6 (Grupo 2) e 7 (Grupo 3).

É possível observar que a média de quantidade de *reads* das amostras HapMap (Grupo 1) está em torno de 98 milhões de *reads* por amostra. Enquanto a média de *reads* das amostras HapMap (Grupo 1) em todos os kits se encontram comparáveis, com todas as médias estando em torno de 3 milhões de diferença (Tabela 5). As amostras do paciente e sua mãe (Grupo 2) feitas pelo SureSelect com 124 milhões de *reads* e NimbleGen com 74 milhões apresentam uma grande diferença entre a quantidade de *reads* (Tabela 6). Ao comparar a cobertura bruta de cada uma das bases, baseando-se somente na quantidade de *reads* e a quantidade de bases nos alvos, tem-se que para as amostras sequenciadas por NimbleGen possuem uma cobertura bruta de 116x e as de SureSelect 245x. As amostras aleatórias (Grupo3) apresentaram uma quantidade de *reads* variadas sendo que para Nextera tem-se em torno de 86 milhões de *reads* (188x), NimbleGen 80 milhões (125x), SureSelect 74 milhões (145x) e SureSelectUTR com 86 milhões de *reads* (115x) (Tabela 7). Apesar da

quantidade de reads ser bastante divergente entre os grupos de comparações, o que pode acarretar um viés de cobertura, em todos os cenários, a média de cobertura bruta atingiu o critério mínimo de 100x.

Tabela 5. Representação da média e mediana do grupo de amostras do HapMap, assim como os números máximos e mínimos dos *reads*, para cada uma das tecnologias.

Tecnologias	Medidas	Nº de reads
Nextera	Média	97.307.852
	Mediana	97.307.852
	Mínimo	97.152.250
	Máximo	97.463.454
NimbleGen	Média	99.959.956
	Mediana	99.959.956
	Mínimo	99.958.084
	Máximo	99.961.828
SureSelect	Média	98.589.533
	Mediana	98.589.533
	Mínimo	98.183.089
	Máximo	98.995.976

Tabela 6. Representação da média e mediana da amostra do paciente e sua mãe, assim como os números máximos e mínimos dos *reads*, para cada uma das tecnologias.

Tecnologias	Medidas	Nº de reads
NimbleGen	Média	74.605.185
	Mediana	74.605.185
	Mínimo	72.560.320
	Máximo	76.650.050
SureSelect	Média	124.000.562
	Mediana	124.000.562
	Mínimo	106.878.349
	Máximo	141.122.775

Tabela 7. Representação da média e mediana do grupo de amostras aleatórias, assim como os números máximos e mínimos dos *reads*, para cada uma das tecnologias.

Tecnologias	Medidas	Nº de reads
Nextera	Média	91.929.866
	Mediana	85.786.383
	Mínimo	65.819.776
	Máximo	146.783.366
NimbleGen	Média	80.845.653
	Mediana	80.413.657
	Mínimo	62.156.142
	Máximo	105.717.594
SureSelect	Média	89.074.800
	Mediana	73.634.604
	Mínimo	39.550.155
	Máximo	244.630.604
SureSelect+UTR	Média	79.975.038
	Mediana	86.156.506
	Mínimo	42.939.150
	Máximo	108.952.739

4.1. Qualidade do sequenciamento

Após o sequenciamento, foi realizada a análise qualitativa dos *reads* pelo FASTQC para cada um dos grupos de amostra⁶⁷. Esta mostrou que a maior parte das bases sequenciadas, em todos os kits de captura, possuíam uma nota Phred ≥ 30 , tendo uma pequena queda na ponta 3' em todos os casos (Figura 7). Os gráficos ilustrativos, encontrados na Figura 8, e a média geral da nota Phred de cada tecnologia, encontradas na Tabela 8, evidenciam uma uniformidade na qualidade dos sequenciamentos. Já o desvio padrão demonstra que as amostras da NimbleGen apresentam maior variabilidade, uma vez que a média da nota por amostra varia de 33 a 35, enquanto as demais tecnologias apresentam variações menores, com mínimo de 33 e máximo de 34 ou de 34 e 35 (Anexo 1).

Figura 8. Gráficos de qualidade gerados pelo programa FASTQC, sendo representativos dos grupos HapMap (A), paciente com hipopituitarismo (B) e amostras aleatórias (C). O eixo y representa a nota Phred com o máximo de 30, e o eixo x a posição da base no *read*, de 1 a 123. Os fundos verde, amarelo e vermelho representam as notas boas, medianas e ruins.

Tabela 8. Média e desvio padrão da nota Phred calculada para cada uma das tecnologias.

Tecnologia	Média	Desvio Padrão
Nextera	34,56	0,5
NimbleGen	34,57	1,4
SureSelect	34	0,47
SureSelect+UTR	34,7	0,5

4.2. Qualidade do mapeamento e taxa de duplicação

Uma vez certificado que as amostras utilizadas apresentavam qualidade de sequenciamento adequada e similares entre si, foi dada a continuidade da análise levando-se em consideração a qualidade do mapeamento. A taxa de mapeamento indica se os fragmentos sequenciados estão dentro da região alvo da tecnologia ou fora (denominado como *off-target*), o que poderia indicar inespecificidade das *probes* de captura. Para um bom mapeamento é importante que os fragmentos tenham um tamanho adequado, uma vez que *reads* menores possuem uma maior probabilidade de se alinharem a múltiplas regiões do genoma. Esse fenômeno diminui a qualidade do mapeamento, podendo levar a um aumento na taxa de *reads* alinhados em regiões *off-target*. Apesar da Nextera apresentar um tamanho de inserto menor do que o esperado pela tecnologia (150pb a 1kb, enquanto a média dos sequenciados foi de 124,3pb)⁸⁷, indicando que a quantidade de bases do *read* referentes ao alvo sejam menores que o tamanho do *read* sequenciado, isso não afetou a qualidade e nem a taxa de mapeamento no alvo, apresentando mediana similar entre as tecnologias NimbleGen e SureSelect+UTR (Tabela 9).

Outra taxa importante para a verificação da qualidade das amostras é a quantidade de *reads* duplicados, uma vez que esses *reads* não podem ser

utilizados para contabilizar cobertura e devem ser retirados antes da chamada de variantes. Uma alta taxa de duplicação pode indicar um viés na captura de região de interesse, potencializada pelos ciclos de PCR no preparo da biblioteca, que podem ter sido excessivos. A tecnologia que apresenta maior mediana de porcentagem de duplicatas é a NimbleGen, que é o dobro da Nextera e até 23% maior que as da Agilent (Tabela 9). Isso é corroborado pela visualização das taxas de duplicação nos *reads* mapeados no genoma e aqueles mapeados dentro da região alvo nos gráficos da Figura 9, que mostra que NimbleGen apresenta mais duplicação em ambos casos, enquanto o SureSelect apresenta maior variabilidade de duplicatas, o que pode indicar um viés técnico no preparo da biblioteca, e o SureSelect+UTR a menor taxa entre todas as tecnologias (Figura 9).

Tabela 9. Média, mediana, mínima e máxima de *reads* mapeados na região total e na região alvo, além da taxa de duplicados por tecnologia, assim como a qualidade do mapeamento.

Tecnologia	Medidas	Porcentagem de reads mapeados no alvo (%)	Porcentagem de reads mapeados no genoma(%)	Tamanho do inserto	Qualidade de mapeamento médio	Porcentagem de reads duplicados totais (%)
Nextera	Média	64,5	99,7	127,1	57,4	12,5
	Mediana	65,8	99,9	116	57,4	13,1
	Mínimo	42,8	98,8	106	57,2	6,8
	Máximo	74,3	99,9	187	57,9	17,6
NimbleGen	Média	67,4	98,9	274,3	56,4	26,9
	Mediana	65,6	98,9	275	56,4	27,9
	Mínimo	61,9	98,3	259	56,4	13,8
	Máximo	76,7	99,5	282	56,5	36,6
SureSelect	Média	78,5	99,9	187	58,8	4,7
	Mediana	78,6	99,9	181,5	59,1	3,5
	Mínimo	76,6	99,8	181	58,5	1,6
	Máximo	80	99,9	196	59,2	10,2
SureSelect+UTR	Média	67	99,8	188,9	59,1	4,7
	Mediana	69,3	99,8	189,5	59,1	4,6
	Mínimo	54,8	99,7	174	59,1	2,5
	Máximo	71,9	99,9	196	59,2	7,7

Figura 9. Gráficos ilustrando a frequência da taxa de duplicação global (acima) e por região (abaixo) nas amostras para cada tecnologia, sendo o eixo x a porcentagem de duplicação e o eixo y a frequência na qual se encontram essas duplicações.

4.3. Profundidade de Cobertura Global

Teoricamente, é esperado que tecnologias que se propõe a cobrir uma região maior do genoma, que seria o caso da SureSelect+UTR (75Mb), tenha uma maior quantidade de *reads*. Apesar da quantidade de *reads* das amostras

do Grupo2 da NimbleGen ser mais baixa, foi o suficiente para atingir o máximo de cobertura da tecnologia, como pode ser visto nos gráficos de curva de saturação (Figuras 9). Entretanto, isso pode acarretar em um viés na cobertura na comparação entre essas duas amostras.

Na análise da quantidade de reads do Grupo 3 de amostras aleatórias foi observado que ela fica em terceiro lugar em relação a sua média de *reads* por amostra, o que ocasionaria uma menor profundidade por região. A partir deste fato, foi realizada uma curva de saturação de *reads* por regiões cobertas de acordo com as profundidades 1x (considerada *drop-off rate*), 10x (mínimo de cobertura esperada) e 20x (corte para chamada de variantes confiável) e foi constatado que a quantidade de *reads* sequenciada para cada uma das tecnologias foi o suficiente para se alcançar o máximo de cobertura da tecnologia. Como pode-se observar nos gráficos da figura 10, a SureSelect+UTR, em vermelho, apresenta o ponto de saturação em torno de 60 milhões de *reads* para profundidade de 10x, cobrindo em média 95% da região alvo. Em contrapartida, a Nextera tem seu ponto de saturação em torno de 70 milhões de *reads* para a mesma profundidade, cobrindo a mesma porcentagem da região alvo, o que indica que essa tecnologia, apesar de cobrir uma menor região genômica (45Mb), necessita de maior quantidade de *reads* para alcançar a mesma profundidade de cobertura do exoma global (Figura 10).

Figura 10. Gráficos de saturação pela quantidade de *reads*, sendo o eixo x os *reads* em milhões e o eixo y a porcentagem da região alvo coberta, no caso a região global, nas profundidades **A)** 1x, **B)** 10x e **C)** 20x.

Figura 11. Relação entre a profundidade de cobertura, até 50x (eixo x), e a porcentagem da região proposta a ser coberta pela tecnologia (eixo y), com uma linha indicando a cobertura ideal de 20x nos grupos **A)** HapMap **B)** paciente de hipopituitarismo e sua mãe e **C)** de amostras aleatórias.

4.4. Análise das *probes* desenhadas para a região de interesse nos diferentes kits

A fim de capturar as regiões de interesse do DNA que será sequenciado, são desenhadas *probes* diferentes para que sejam selecionados os fragmentos de interesse. Cada uma das tecnologias utilizadas possui *probes* diferentes para essa captura, mesmo que suas finalidades sejam a mesma de sequenciar toda a região exônica do genoma, resultando em diferentes regiões abordadas. As regiões alvo referentes a cada um dos kits de captura estão disponibilizadas por cada empresa em seu website (Nextera⁸⁸; NimbleGen⁸⁹; SureSelect e SureSelect+UTR⁹⁰), e suas diferenças podem acarretar um viés na cobertura de alguns genes devido sua proximidade à *probe*. Dentre as quatro tecnologias utilizadas, a única que possui *probes* desenhadas para a região de interesse representando os 76 genes e que equivale a 161.022 pares de bases, foi a Nextera, seguido pelo NimbleGen, e em terceiro e quarto lugar as tecnologias da Agilent, SureSelect e SureSelect+UTR, respectivamente. Diante disso, concluímos que as 4 diferentes tecnologias possuem 128.300 pares de bases em comum. Figura 12.

Figura 12. Diagrama de Venn representando a intersecção das sondas desenhadas em relação à região dos 76 genes.

4.5. Profundidade de cobertura na região alvo

Após verificar que todas as amostras selecionadas são comparáveis em termo de qualidade, foi iniciada a comparação da profundidade de cobertura. Os genes estudados perfazem 161.022 kb e para cada uma das tecnologias utilizadas foi calculada a profundidade média da região total capturada, assim como a de interesse, para cada um dos grupos. Como pode ser visto no gráfico (Figura 13B e 13C), o kit de captura da NimbleGen possui a menor cobertura do alvo conforme o aumento de profundidade. Quanto maior a exigência de profundidade, menor a região coberta pelos *reads*. Isso pode ser parcialmente explicado pela baixa quantidade de *reads* sequenciados em relação as outras tecnologias que abordam alvos menores que a NimbleGen (de 64Mb), além de ter apresentado a menor taxa de mapeamento no alvo de 65.6%. Um outro fator que pode demonstrar um viés de cobertura por vir do preparo da

biblioteca, uma vez que o mesmo padrão não é visto no gráfico que representa as amostras do HapMap. A tecnologia que melhor reflete esse viés para este parâmetro foi o kit de captura do Nextera que apresenta menor rendimento nas coberturas iniciais, indicando que existem regiões com viés. Isso pode ser mais claramente observado na figura 11C com 20x de cobertura, que mostra que apesar do aumento da quantidade de *reads* sequenciados, não é possível observar um aumento da porcentagem coberta da região alvo, sugerindo que existam regiões com viés de captura por essa tecnologia. Já as tecnologias da SureSelect estão acima das demais nas profundidades, sendo constantemente a que atinge maior cobertura em todos os casos. A boa representatividade das tecnologias SureSelect e Nextera podem se dar devido ao fato de ambas tecnologias possuírem região alvo menor do que a SureSelect+UTR (75Mb) e NimbleGen (64Mb). Pode-se também verificar que esses comportamentos dentro da região alvo não são diferentes daqueles encontrados nas regiões totais, como é comparável nos gráficos das figuras 13 e 14. Entretanto, quando se trata da tecnologia da NimbleGen, a cobertura em 20x para a região alvo sofre uma queda de 2 a 4% na cobertura em relação a cobertura geral nesse mesmo corte, padrão esse encontrado em todos os três grupos de amostra testados, como visto nas Figuras 13 e 14. A SureSelect, inclusive, mantém sua taxa de cobertura (96 a 98%) em todos os grupos, quando se compara a cobertura da região total e da região de interesse.

Figura 13. Gráficos de saturação pela quantidade de *reads*, sendo o eixo x os *reads* em milhões e o eixo y a porcentagem da região alvo coberta, no caso a região de interesse dos 76 genes, nas profundidades **A)** 1x, **B)** 10x e **C)** 20x.

Figura 14. Relação entre a profundidade de cobertura, até 50x (eixo x), e a porcentagem da região de 76 genes de interesse (eixo y), com uma linha indicando a cobertura ideal de 20x nos grupos **A)** HapMap **B)** paciente de hipopituitarismo e sua mãe e **C)** de amostras aleatórias.

4.6. Cobertura das regiões de acordo com a quantidade de GC

Uma das maiores dificuldades em sequenciamentos, seja pelas metodologias atuais ou a de Sanger^{45,91}, são regiões muito ricas (aquelas que possuem >80% de GC) ou pobres (<20%) em GC. Portanto, foi verificado também a relação entre a cobertura da região considerando a porcentagem de GC presente nessas regiões (Figuras 15 e 16). Todas as tecnologias possuem um viés contra essas altas e baixas taxas de GC, não tendo cobertura nessas regiões extremas, mas é notável a diferença no alcance de cobertura da tecnologia SureSelect em relação as demais tanto nas amostras do HapMap quanto nas do terceiro grupo de amostras, taxa que explicaria uma maior profundidade na cobertura em regiões com GC em relação as demais. Tanto a SureSelect quanto a Nextera tem maior capacidade de cobrir essas regiões de forma uniforme em relação a tecnologia da NimbleGen.

Figura 15. Gráficos de densidade da quantidade de GC por tecnologia. Observa-se que a tecnologia NimbleGen possui a menor capacidade de cobertura das regiões ricas em GC no grupo HapMap.

Figura 16. Gráficos de densidade da quantidade de GC por tecnologia. Observa-se que a tecnologia NimbleGen possui a menor capacidade de cobertura das regiões ricas em GC no grupo de amostras aleatorias.

Uma comparação entre o grupo de 76 genes de interesse e 76 genes aleatórios do genoma mostra que, apesar de não ser estatisticamente significativo de acordo com o teste de Kolmogorov-Smirnov, há uma diferença entre a quantidade de GC entre os dois grupos. Como pode ser visto na Figura 17, os 76 genes de interesse apresentam uma maior frequência de regiões com uma taxa de GC entre 65 e 75%, assim como nas taxas de 45 a 60%.

Mostrando que esse pode ser uma fonte de viés na cobertura em algumas das tecnologias, como no caso da NimbleGen, que não apresenta a mesma capacidade que as demais de cobrir essas regiões de forma uniforme.

Figura 17. Gráfico de comparação entre os 76 genes selecionados para esse estudo (em azul) e 76 genes aleatórios do genoma (em rosa). O eixo X é a frequência e o eixo Y é a quantidade de GC entre 30 e 80%.

4.7. Cobertura dos genes

A análise de cobertura de cada gene foi feita principalmente utilizando o programa coverageBed, que retornou a cobertura por base de todos os exons do gene, ou seja, o número de *reads* diferentes que continham a mesma base. Para verificar o quanto da extensão de genes foi coberto, foram contadas quantas bases do gene foram cobertas acima de 1x (*drop-off rate*), 10x (mínimo) e 20x (cobertura para chamada de variantes confiável) a fim de descobrir quanto da extensão do gene foi coberta nessas faixas.

Foram então separados os genes que possuísem 100% de cobertura em sua extensão em cada uma dessas faixas de reads, com a quantidade de genes com essa porcentagem de cobertura fosse inversamente proporcional ao aumento da profundidade de cobertura. (Tabela 10 a 12/Anexo 2).

Tabela 10. Quantidade de genes com 100% de cobertura de extensão nas amostras HapMap

Tecnologia	Cobertura 1x	Cobertura 10x	Cobertura 20x
Nextera	70	51	34
NimbleGen	55	43	28
SureSelect	73	61	37

Tabela 11. Quantidade de genes com 100% de cobertura de extensão nas amostras do paciente com hipopituitarismo

Tecnologia	Cobertura 1x	Cobertura 10x	Cobertura 20x
NimbleGen	62	43	22
SureSelect	74	69	57

Tabela 12. Quantidade de genes com 100% de cobertura de extensão em cada uma das tecnologias.

Tecnologia	Cobertura 1x	Cobertura 10x	Cobertura 20x
Nextera	64	20	11
NimbleGen	66	47	25
SureSelect	74	52	26
SureSelect+UTR	74	30	15

Foram verificados também quais foram os 5 genes que apresentavam as piores coberturas dentre todos da mesma tecnologia com os mesmos cortes de reads (Anexo 3). O gene com pior cobertura em todos os cortes foi o *HES5*, tendo uma cobertura de 86,2% na Nextera em 1x, 43,1% e 34,3% pela NimbleGen em 10x e 20x, respectivamente (Tabela 13). Esse gene é um gene rico em GC (70% de GC), tendo em seus três exons uma média de quantidade de GC de 66,6%, e sua primeira porção codificadora possui 76,8% de GC,

sendo a que possui pior média de cobertura principalmente nas tecnologias Nextera e NimbleGen, enquanto as tecnologias da SureSelect apresentam melhor desempenho nessas mesmas regiões (Figura 17), indo de acordo com o que pode ser observado sobre a capacidade de cobertura das regiões ricas em GC no último tópico⁹². Apesar desse gene não ter variantes descritas como causadoras do fenótipo de interesse, esse gene trabalha em conjunto com o *HES1* para o correto desenvolvimento da glândula pituitária como um todo, em especial a diferenciação entre os lobo intermediário e anterior. O camundongo com deleção em ambos genes não apresentavam o lobo posterior⁹³.

Tabela 13. Tabela mostrando a extensão da cobertura do gene *HES5* em cada uma das tecnologias.

Tecnologia	Cobertura 1x (%)	Cobertura 10x (%)	Cobertura 20x (%)
Nextera	86,2	69,4	58,3
NimbleGen	93,4	43,1	34,3
SureSelect	99,1	89,9	85,8
SureSelect+UTR	98,5	86,7	76,8

Figura 18. Gráficos apresentando a média de cobertura de cada uma das tecnologias para o gene *HES5*. O eixo y é a quantidade de reads de 0 a 100 e o eixo x mostra a extensão do gene, cada duas marcações sendo o início e o final do exon. Seu primeiro exon, que está entre as duas primeiras marcações no eixo x, possui 76,8% de quantidade de GC, que é muito próximo a capacidade máxima de cobertura de todas as tecnologias. A tecnologia SureSelect apresenta um pico alto de cobertura que ultrapassa os 100x no segundo exon.

Entre os genes relacionados ao hipopituitarismo, o que possui pior cobertura entre todas as tecnologias foi o *SOX3*. Esse gene possui apenas um exon que contém 70% de GC. Apesar da tecnologia da NimbleGen apenas ser capaz de cobrir 92,1%, 61,1% e 43,4% do gene com 1x, 10x e 20x de cobertura respectivamente (Tabela 16/Figura 14), é um gene bem representado na tecnologia da SureSelect, com apenas o início do gene apresentando uma cobertura abaixo de 20x, corroborando novamente a teoria de que essa tecnologia é capaz de atingir um nível maior de cobertura de profundidade nessas regiões. Estudos em camundongos com deleção nesse gene apresentaram um desenvolvimento anormal da bolsa de Rathke, com os

defeitos pituitários e hipotalâmicos presentes no período pós natal. A duplicação nesse gene também foram relacionadas ao hipopituitarismo^{39,40,94}.

Tabela 14. Extensão de cobertura do gene SOX3 em todas as quatro tecnologias.

Tecnologia	Cobertura 1x (%)	Cobertura 10x (%)	Cobertura 20x (%)
Nextera	99,5	65,6	55,3
NimbleGen	99,1	31,7	7
SureSelect	100	88,9	74,6
SureSelect+UTR	100	83,8	63

Figura 19. Gráficos apresentando a média de cobertura do gene SOX3 em todas as tecnologias. O eixo y representa a quantidade de reads, de 0 a 100x, e o x a extensão do gene, que possui apenas um exon. A tecnologia Nextera apresenta dois picos que se encontram acima de 100x, enquanto a SureSelect apresenta apenas um. É possível também visualizar que as médias de cobertura das tecnologias da SureSelect seguem o mesmo padrão.

Ambos os genes SOX3 e HES5 possuem uma porcentagem de GC alta, estando localizados acima do percentil 90 (0,67), evidenciando que ambos os genes possuem porcentagem de GC maior que 90% entre os 76 genes estudados (figura 15).

Figura 20. Histograma da porcentagem de GC em relação aos 69 genes. O eixo x representa a porcentagem de GC e o eixo y a quantidade de genes dentro de cada um dos intervalos. Os genes *HES5* (média de 68%) e *SOX3* (70%) estão no gráfico em azul e preto, respectivamente.

4.8. Variantes não descritas em bancos de dados nos genes de interesse

Partindo das variantes encontradas no banco de dados ABraOM, foi verificado que 8.542 variantes se encontravam na nossa região de interesse. O programa ANNOVAR foi utilizado para anotar as variantes encontradas apenas nos 76 genes estudados. Uma filtragem semelhante aquela feita no ABraOM porém contemplando apenas as VCF dos genes de interesse e retirando qualquer variante já descrita em bancos de dados, evidenciou 885 variantes novas nessa região (Figura 16), ou seja, exclusivas da população brasileira⁸⁵.

Figura 21. Filtragem feita nas 8.542 variantes do ABraOM a fim de descobrir quantas seriam exclusivas da população estudada.

5. Discussão

5. Discussão

Na busca de variantes alélicas a partir dos métodos de sequenciamento de larga escala, se faz necessário uma análise prévia das métricas que possa demonstrar uma taxa de cobertura mínima exigida para que não se percam aquelas que possam ser classificadas como patogênicas. De um modo geral, essas informações sobre a cobertura dos genes associados ao hipopituitarismo encontram-se ausentes nos artigos recentemente publicados. Uma análise inicial das amostras dos pacientes com hipopituitarismo usando o kit da NimbleGen identificou uma cobertura ruim principalmente em genes ricos em GC. Diante disso, o escopo do presente estudo foi comparar a eficiência de cobertura de 4 kits disponíveis no mercado e acessíveis a nós. Outros estudos comparando pelo menos 3 dos kits a nós disponíveis já tinham sido feitos^{15,32,65}, com isso optamos por expandir a análise das métricas nas amostras brasileiras e na população de pacientes brasileiros com hipopituitarismo. Com esse propósito foi incluída a análise de uma região que contivesse os genes relacionados ao hipopituitarismo e/ou papel na embriogênese hipofisária a fim de se considerar o melhor kit a ser utilizado em coortes de pacientes com essa mesma doença.

O uso de outras tecnologias mais simples que garantem melhor cobertura, como é o caso do painel de genes, pode parecer tentadora quando se trata de genes de difícil sequenciamento como os analisados nesse estudo. Até hoje, poucos estudos foram realizados utilizando essa técnica em coortes de hipopituitarismo, mas os estudos já realizados apresentam uma baixa taxa de diagnóstico molecular, como é o caso de Pérez Millan e colaboradores⁹⁵, que fizeram o sequenciamento de 51 pacientes e encontraram apenas uma variante

no gene GH1, que é um gene prevalente nessa doença. Isso, juntamente com os levantamentos na literatura feitos por DeRienzo³⁸ e Fang⁹, que apontam a baixa taxa de diagnóstico molecular com os genes já descritos na literatura, levam pesquisadores a utilizarem a tecnologia que permita buscar variantes em uma maior região do genoma como a melhor opção inicial.

O resultado encontrado no presente estudo vai de acordo com resultados obtidos em outros estudos de comparação feitos com exoma^{15,32,65,96,97}, onde o SureSelect foi apontado como o melhor kit em questão de eficiência e cobertura. Entretanto, todos esses estudos utilizaram de uma a três amostras para sua análise. Logo, o diferencial do presente estudo foi utilizar amostras provindas de uma coorte brasileira, que é uma população pouco estudada em comparação a outras, algo que pode ser comprovado pela sua ausência nos maiores bancos de dados internacionais de frequência de variantes.

No presente estudo, os três grupos analisados (1 - HapMap, 2 - paciente com hipopituitarismo e 3 - amostras brasileiras randômicas) apresentaram boa qualidade de sequenciamento, com número de reads sequenciados, taxa de qualidade de mapeamento e duplicação comparáveis, permitindo que a análise do resultado final se mantivesse semelhante nos 3 diferentes grupos, concluindo que diferentes amostras realizadas em diferentes kits não foi um fator limitante na conclusão da capacidade de cobertura dos kits utilizados para preparação de biblioteca para sequenciamento em larga escala.

Foi feita também uma comparação entre a cobertura do exoma global com a cobertura atingida em uma região de interesse, de 161.022 pares de bases, que continha todos os genes já descritos relacionados ao hipopituitarismo ou que sabidamente tem seu papel na embriogênese

hipofisária. De acordo com os resultados obtidos, o kit da NimbleGen teve uma pequena queda na porcentagem de cobertura nessa região em relação a sua capacidade de cobertura no exoma global, algo que foi visto em todos os três grupos de amostra. Isso pode ser devido a quantidade de GC presente na região de interesse, que apesar de não ser estatisticamente significativa, pode influenciar o resultado desse kit devido seu viés em relação a essas regiões. Outras comparações feitas não singularizam o kit da NimbleGen como tendo dificuldades nessas regiões, apontando apenas uma queda na sua cobertura, mas entram em conflito sobre a capacidade dos kits da Illumina (aqui apenas o Nextera) de cobrir essas regiões, por hora tendo a cobertura mais uniforme entre as demais¹⁵ e por outra tendo a mesma queda de cobertura que a NimbleGen⁶⁵. Neste caso, a melhor tecnologia seria a da SureSelect, que não perde cobertura entre ambas as regiões, assim como mostra uma cobertura alta e uniforme nas regiões com GC.

Apesar da cobertura ideal para a chamada de variantes ser de 20x, ou seja, a mesma base ser coberta em 20 *reads* diferentes, é importante iniciar nossa análise de cobertura horizontal dos genes com o *drop-off rate* de 1x, uma vez que isso indica se a metodologia é capaz de cobrir completamente o gene. Infelizmente, em nenhum dos grupos analisados, atingiu-se 100% de cobertura horizontal em todos os 76 genes. O número de genes não completamente cobertos cai à medida que se aumenta a cobertura vertical para 10x e 20x, chegando a 11 genes cobertos pelo Nextera nessas condições, sendo o pior resultado entre todos os kits. Foi feita então a análise específica do gene que tinha a pior cobertura em todas as metodologias, no caso o *HES5*. Esse gene, apesar de ter seu papel na embriogênese hipofisária, não foi

associado à mutação no humano levando a quadro de hipopituitarismo. Por isso, foi escolhido também o gene com pior cobertura dentre esses que estão associados à doença, e foi feita também a análise específica do gene *SOX3*. Ambos genes são ricos em GC, tendo uma média de 68 a 70% de GC em seus exons. Hipotetizamos então que essa é a causa para a má cobertura desses genes. A cobertura do gene *SOX3* é especialmente importante, uma vez que é um gene classicamente envolvido em quadros de hipopituitarismo, tendo diversas variantes descritas nele. Por isso, é de suma importância que se tenha certeza que esse gene foi coberto no exoma, para poder averiguar a presença ou ausência de variantes alélicas patogênicas nesse gene com confiança.

A utilização de amostras brasileiras nos levou também a analisar quantas das variantes presentes no banco de dados brasileiro ABraOM⁸⁵ estavam contidas na região estudada de interesse. Dentre as 207.621 variantes específicas, de acordo com a ausência em bancos de dados internacionais, 885 se encontravam em genes relacionados ao hipopituitarismo e/ou presentes na embriogênese hipofisária. Isso evidencia novamente o quão importante é termos estudos específicos para nossa população, que ainda tem seu material genética amplamente desconhecido.

Por fim, entre os quatro kits analisados, o com a menor eficiência de cobertura foi o kit da Nextera, que apresentou uma cobertura menor do que as demais tecnologias mesmo tendo tido número suficiente, assim como um viés para algumas regiões, como visto nos gráficos de cobertura dos genes. Isso pode ser explicado pelo tipo de fragmentação utilizada em seu protocolo, chamada tagmentação. Esse método, por não ser randômico como a metodologia mecânica, apresenta viés de cobertura por ter afinidade por

determinados fragmentos. Entretanto, a Illumina tem lançado outros kits que permitem optar pela a utilização da fragmentação mecânica, e poderiam apresentar melhores resultados do que o kit Nextera aqui analisado.

6. Conclusão

6. Conclusão

O presente estudo se dispôs a analisar a cobertura dos genes envolvidos na embriogênese hipofisária e relacionados ao hipopituitarismo congênito em humanos, a partir de busca na literatura, utilizando quatro diferentes kits de preparação de biblioteca em três grupos distintos (1 - HapMap, 2 - Paciente com hipopituitarismo e sua mãe, 3 - amostra aleatórias da população brasileira).

A partir da análise sistemática em três grupos distintos de amostras verificou-se a qualidade do sequenciamento das diversas tecnologias e concluiu-se que o kit de preparação de biblioteca SureSelect, da Agilent apresenta a melhor eficiência e taxa de cobertura tanto regiões globais e, principalmente, para as regiões de interesse para esse estudo. Esta observação é importante, uma vez que o diagnóstico molecular para o hipopituitarismo ainda é pequeno. Assim utilizar a melhor tecnologia que aborda os alvos de interesse é relevante no contexto da doença. Analisando as variantes da população brasileira, utilizando o banco de dados ABraOM comparado com outros bancos internacionais, concluimos que a população brasileira possui 885 variantes alélicas exclusivas comparadas aos bancos internacionais, mostrando uma particularidade dessa população.

7. Referências

7. Referências

- 1 Zhu X, Gleiberman AS, Rosenfeld MG. Molecular Physiology of Pituitary Development: Signaling and Transcriptional Networks. *Physiol Rev* 2007;87:933–63. doi:10.1152/physrev.00006.2006.
- 2 Antunes-Rodrigues J, LLK E, M DC. Neuroendocrinologia básica e aplicada. Rio de Janeiro: Guanabara Koogan; 2005.
- 3 Rizzoti K. Genetic regulation of murine pituitary development. *J Mol Endocrinol* 2015;54:R55–73. doi:10.1530/JME-14-0237.
- 4 Takuma N, Sheng HZ, Furuta Y, Ward JM, Sharma K, Hogan BL, et al. Formation of Rathke's pouch requires dual induction from the diencephalon. *Development* 1998;125:4835–40.
- 5 Dasen JS, Rosenfeld MG. Signaling and Transcriptional Mechanisms in Pituitary Development. *Annu Rev Neurosci* 2001;24:327–55. doi:10.1146/annurev.neuro.24.1.327.
- 6 Alatzoglou KS, Dattani MT. Genetic forms of hypopituitarism and their manifestation in the neonatal period. *Early Hum Dev* 2009;85:705–12. doi:10.1016/j.earlhumdev.2009.08.057.
- 7 Pfäffle R, Klammt J. Pituitary transcription factors in the aetiology of combined pituitary hormone deficiency. *Best Pract Res Clin Endocrinol Metab* 2011;25:43–60. doi:10.1016/j.beem.2010.10.014.
- 8 Medicine® USNL of. Genetics Home Reference 2007:NF2.
- 9 Fang Q, George AS, Brinkmeier ML, Mortensen AH, Gergics P, Cheung LYM, et al. Genetics of Combined Pituitary Hormone Deficiency:

- Roadmap into the Genome Era. *Endocr Rev* 2016;37:636–75.
doi:10.1210/er.2016-1101.
- 10 Sun Y, Bak B, Schoenmakers N, van Trotsenburg ASP, Oostdijk W, Voshol P, et al. Loss-of-function mutations in IGSF1 cause an X-linked syndrome of central hypothyroidism and testicular enlargement. *Nat Genet* 2012;44:1375–81. doi:10.1038/ng.2453.
 - 11 Schoenmakers N, Alatzoglou KS, Chatterjee VK, Dattani MT. Recent advances in central congenital hypothyroidism. *J Endocrinol* 2015;227:R51–71. doi:10.1530/JOE-15-0341.
 - 12 Rizzo JM, Buck MJ. Key Principles and Clinical Applications of “Next-Generation” DNA Sequencing. *Cancer Prev Res* 2012;5:887–900. doi:10.1158/1940-6207.CAPR-11-0432.
 - 13 Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;74:5463–7.
 - 14 Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet* 2010;11:31–46. doi:10.1038/nrg2626.
 - 15 Chilamakuri CS, Lorenz S, Madoui M-A, Vodák D, Sun J, Hovig E, et al. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014;15:449. doi:10.1186/1471-2164-15-449.
 - 16 Sastre L. Exome sequencing: what clinicians need to know. *Adv Genomics Genet* 2014;4:15. doi:10.2147/AGG.S39108.
 - 17 Kiezun A, Garimella K, Do R, Stitzel NO, Neale BM, McLaren PJ, et al.

- Exome sequencing and the genetic basis of complex traits. *Nat Genet* 2012;44:623–30. doi:10.1038/ng.2303.
- 18 Wang Z, Liu X, Yang B-Z, Gelernter J. The Role and Challenges of Exome Sequencing in Studies of Human Diseases. *Front Genet* 2013;4:160. doi:10.3389/fgene.2013.00160.
- 19 Daber R, Sukhadia S, Morrisette JJD. Understanding the limitations of next generation sequencing informatics, an approach to clinical pipeline validation using artificial data sets. *Cancer Genet* 2013;206:441–8. doi:10.1016/j.cancergen.2013.11.005.
- 20 Ellingford JM, Campbell C, Barton S, Bhaskar S, Gupta S, Taylor RL, et al. Validation of copy number variation analysis for next-generation sequencing diagnostics. *Eur J Hum Genet* 2017;25:719–24. doi:10.1038/ejhg.2017.42.
- 21 Day-Williams AG, Zeggini E. The effect of next-generation sequencing technology on complex trait research. *Eur J Clin Invest* 2011;41:561–7. doi:10.1111/j.1365-2362.2010.02437.x.
- 22 Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? *J Med Genet* 2011;48:580–9. doi:10.1136/jmedgenet-2011-100223.
- 23 Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* 2010;19:R145–51. doi:10.1093/hmg/ddq333.
- 24 Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 2008;92:255–64.

- doi:10.1016/j.ygeno.2008.07.001.
- 25 Lin X, Tang W, Ahmad S, Lu J, Colby CC, Zhu J, et al. Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. *Hear Res* 2012;288:67–76. doi:10.1016/j.heares.2012.01.004.
 - 26 Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci* 2009;106:19096–101. doi:10.1073/pnas.0910672106.
 - 27 Antonarakis S, Krawczak M, Cooper D. The nature and mechanisms of human gene mutation. *Online Metab. Mol. Bases Inherit. Dis.*, 2012.
 - 28 Zhang X. Exome sequencing greatly expedites the progressive research of Mendelian diseases. *Front Med* 2014;8:42–57. doi:10.1007/s11684-014-0303-9.
 - 29 Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: Overviews and challenges. *Biotechniques* 2014;56:61–4, 66, 68, passim. doi:10.2144/000114133.
 - 30 Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med* 2013;15:733–47. doi:10.1038/gim.2013.92.
 - 31 Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery.

- Nat Rev Genet 2011;12:745–55. doi:10.1038/nrg3031.
- 32 Shigemizu D, Momozawa Y, Abe T, Morizono T, Boroevich KA, Takata S, et al. Performance comparison of four commercial human whole-exome capture platforms. Sci Rep 2015;5:12742. doi:10.1038/srep12742.
 - 33 Goodwin S, McPherson JD, McCombie WR. No Title 2016;17. doi:10.1038/nrg.2016.49.
 - 34 Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 2012;131:1541–54. doi:10.1007/s00439-012-1213-z.
 - 35 D’Antonio M, D’Onorio De Meo P, Paoletti D, Elmi B, Pallocca M, Sanna N, et al. No Title 2013:S11. doi:10.1186/1471-2105-14-S7-S11.
 - 36 Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011;12:443–51. doi:10.1038/nrg2986.
 - 37 Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics 2012;13 Suppl 8:S8. doi:10.1186/1471-2164-13-S8-S8.
 - 38 De Rienzo F, Mellone S, Bellone S, Babu D, Fusco I, Prodam F, et al. Frequency of genetic defects in combined pituitary hormone deficiency: a systematic review and analysis of a multicentre Italian cohort. Clin Endocrinol (Oxf) 2015;83:849–60. doi:10.1111/cen.12849.
 - 39 Woods KS, Cundall M, Turton J, Rizotti K, Mehta A, Palmer R, et al. Over- and Underdosage of SOX3 Is Associated with Infundibular

- Hypoplasia and Hypopituitarism. *Am J Hum Genet* 2005;76:833–49.
doi:10.1086/430134.
- 40 Laumonnier F, Ronce N, Hamel BCJ, Thomas P, Lespinasse J, Raynaud M, et al. Transcription Factor SOX3 Is Involved in X-Linked Mental Retardation with Growth Hormone Deficiency. *Am J Hum Genet* 2002;71:1450–5. doi:10.1086/344661.
- 41 Kelberman D, Rizzoti K, Avilion A, Bitner-Glindzicz M, Cianfarani S, Collins J, et al. Mutations within Sox2/SOX2 are associated with abnormalities in the hypothalamo-pituitary-gonadal axis in mice and humans. *J Clin Invest* 2006;116:2442–55. doi:10.1172/JCI28658.
- 42 Numakura C, Kitanaka S, Kato M, Ishikawa S, Hamamoto Y, Katsushima Y, et al. Supernumerary impacted teeth in a patient with SOX2 anophthalmia syndrome. *Am J Med Genet Part A* 2010;152A:2355–9. doi:10.1002/ajmg.a.33556.
- 43 Alatzoglou KS, Andoniadou CL, Kelberman D, Buchanan CR, Crolla J, Arriazu MC, et al. SOX2 haploinsufficiency is associated with slow progressing hypothalamo-pituitary tumours. *Hum Mutat* 2011;32:1376–80. doi:10.1002/humu.21606.
- 44 Stark Z, Storen R, Bennetts B, Savarirayan R, Jamieson R V. Isolated hypogonadotropic hypogonadism with SOX2 mutation and anophthalmia/microphthalmia in offspring. *Eur J Hum Genet* 2011;19:753–6. doi:10.1038/ejhg.2011.11.
- 45 Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS*

- One 2013;8:e62856. doi:10.1371/journal.pone.0062856.
- 46 Webb EA, AlMutair A, Kelberman D, Bacchelli C, Chanudet E, Lescai F, et al. ARNT2 mutation causes hypopituitarism, post-natal microcephaly, visual and renal anomalies. *Brain* 2013;136:3096–105. doi:10.1093/brain/awt218.
 - 47 Smith JDD, Hing AV V, Clarke CMM, Johnson NMM, Perez FAA, Park SSS, et al. No Title 2014;95. doi:10.1016/j.ajhg.2014.07.008.
 - 48 Synofzik M, Gonzalez MA, Lourenco CM, Coutelier M, Haack TB, Rebelo A, et al. PNPLA6 mutations cause Boucher-Neuhauser and Gordon Holmes syndromes as part of a broad neurodegenerative spectrum. *Brain* 2014;137:69–77. doi:10.1093/brain/awt326.
 - 49 Hufnagel RB, Arno G, Hein ND, Hersheson J, Prasad M, Anderson Y, et al. Neuropathy target esterase impairments cause Oliver-McFarlane and Laurence-Moon syndromes. *J Med Genet* 2015;52:85–94. doi:10.1136/jmedgenet-2014-102856.
 - 50 Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu Y-F, McSweeney KM, et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med* 2015;17:774–81. doi:10.1038/gim.2014.191.
 - 51 Karaca E, Buyukkaya R, Pehlivan D, Charng W-L, Yaykasli KO, Bayram Y, et al. Whole-exome sequencing identifies homozygous GPR161 mutation in a family with pituitary stalk interruption syndrome. *J Clin Endocrinol Metab* 2015;100:E140-7. doi:10.1210/jc.2014-1984.
 - 52 Bashamboo A, Bignon-Topalovic J, Rouba H, McElreavey K, Brauner R.

- A Nonsense Mutation in the Hedgehog Receptor CDON Associated With Pituitary Stalk Interruption Syndrome. *J Clin Endocrinol Metab* 2016;101:12–5. doi:10.1210/jc.2015-2995.
- 53 Gregory LC, Gevers EF, Baker J, Kasia T, Chong K, Josifova DJ, et al. Structural pituitary abnormalities associated with CHARGE syndrome. *J Clin Endocrinol Metab* 2013;98:E737-43. doi:10.1210/jc.2012-3467.
- 54 Lucas-Herald AK, Kinning E, Iida A, Wang Z, Miyake N, Ikegawa S, et al. A Case of Functional Growth Hormone Deficiency and Early Growth Retardation in a Child With IFT172 Mutations. *J Clin Endocrinol Metab* 2015;100:1221–4. doi:10.1210/jc.2014-3852.
- 55 Tata B, Huijbregts L, Jacquier S, Csaba Z, Genin E, Meyer V, et al. Haploinsufficiency of Dmxi2, encoding a synaptic protein, causes infertility associated with a loss of GnRH neurons in mouse. *PLoS Biol* 2014;12:e1001952. doi:10.1371/journal.pbio.1001952.
- 56 Takagi M, Narumi S, Hamada R, Hasegawa Y, Hasegawa T. A novel KAL1 mutation is associated with combined pituitary hormone deficiency. *Hum Genome Var* 2014;1:14011. doi:10.1038/hgv.2014.11.
- 57 Tommiska J, Käsäkoski J, Skibsbye L, Vaaralahti K, Liu X, Lodge EJ, et al. Two missense mutations in KCNQ1 cause pituitary hormone deficiency and maternally inherited gingival fibromatosis. *Nat Commun* 2017;8:1289. doi:10.1038/s41467-017-01429-z.
- 58 Starink E, Hokken-Koelega ACS, Visser TJ, Baan J, Peeters RP, de Graaff LCG. Genetic analysis of IRF6, a gene involved in craniofacial midline formation, in relation to pituitary and facial morphology of patients

- with idiopathic growth hormone deficiency. *Pituitary* 2017;20:499–508. doi:10.1007/s11102-017-0808-8.
- 59 Lal RA, Bachrach LK, Hoffman AR, Inlora J, Rego S, Snyder MP, et al. A Case Report of Hypoglycemia and Hypogammaglobulinemia: DAVID Syndrome in a Patient With a Novel NFKB2 Mutation. *J Clin Endocrinol Metab* 2017;102:2127–30. doi:10.1210/jc.2017-00341.
- 60 Giri D, Vignola ML, Gualtieri A, Scagliotti V, McNamara P, Peak M, et al. Novel FOXA2 mutation causes Hyperinsulinism, Hypopituitarism with Craniofacial and Endoderm-derived organ abnormalities. *Hum Mol Genet* 2017;26:4315–26. doi:10.1093/hmg/ddx318.
- 61 Thakur M, Taha D, Misra VK. A Case of Congenital Hypopituitarism Associated With a 1p31 Microdeletion: A Possible Role for LEPR and JAK1. *J Endocr Soc* 2017;1:278–82. doi:10.1210/js.2016-1072.
- 62 Simm F, Griesbeck A, Choukair D, Weiß B, Paramasivam N, Klammt J, et al. Identification of SLC20A1 and SLC15A4 among other genes as potential risk factors for combined pituitary hormone deficiency. *Genet Med* 2018;20:728–36. doi:10.1038/gim.2017.165.
- 63 D Hidalgo-Santos A, Del Carmen DeMingo-Aleman M, Moreno-Macián F, Roselló M, Orellana C, Martínez F, et al. A Novel Mutation of MAGEL2 in a Patient with Schaaf-Yang Syndrome and Hypopituitarism. *Int J Endocrinol Metab* 2018;16:e67329. doi:10.5812/ijem.67329.
- 64 Zhang G, Wang J, Yang J, Li W, Deng Y, Li J, et al. Comparison and evaluation of two exome capture kits and sequencing platforms for variant calling. *BMC Genomics* 2015;16:581. doi:10.1186/s12864-015-1796-6.

- 65 Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* 2011;29:908–14. doi:10.1038/nbt.1975.
- 66 Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988;16:1215.
- 67 Andrews S. FastQC: a quality control tool for high throughput sequence data 2010.
- 68 Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998;8:186–94.
- 69 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–60. doi:10.1093/bioinformatics/btp324.
- 70 Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res* 2002;12:996–1006. doi:10.1101/gr.229102. Article published online before print in May 2002.
- 71 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2. doi:10.1093/bioinformatics/btq033.
- 72 García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* 2012;28:2678–9. doi:10.1093/bioinformatics/bts503.

- 73 Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2015;32:btv566. doi:10.1093/bioinformatics/btv566.
- 74 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
- 75 DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491–8. doi:10.1038/ng.806.
- 76 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–303. doi:10.1101/gr.107524.110.
- 77 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.*, vol. 43, Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013, p. 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43.
- 78 Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010;467:1061–73. doi:10.1038/nature09534.
- 79 OMIM - Online Mendelian Inheritance in Man n.d. <https://www.omim.org/> (accessed July 24, 2017).

- 80 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–11.
- 81 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164–e164. doi:10.1093/nar/gkq603.
- 82 ExAC project pins down rare gene variants. *Nature* 2016;536:249–249. doi:10.1038/536249a.
- 83 Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res* 2017;45:D840–5. doi:10.1093/nar/gkw971.
- 84 Exome Variant Server n.d. <http://evs.gs.washington.edu/EVS/> (accessed July 24, 2017).
- 85 Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat* 2017;38:751–63. doi:10.1002/humu.23220.
- 86 Pena SDJ, Bastos-Rodrigues L, Pimenta JR, Bydlowski SP. DNA tests probe the genomic ancestry of Brazilians. *Brazilian J Med Biol Res = Rev Bras Pesqui Medicas e Biol* 2009;42:870–6.
- 87 Illumina. FOR RESEARCH USE ONLY Nextera ® Rapid Capture Guide 2013.

- 88 Nextera Rapid Capture Exome v1.2 Product Files n.d.
<https://support.illumina.com/downloads/nextera-rapid-capture-exome-v1-2-product-files.html> (accessed July 24, 2017).

- 89 SeqCap EZ Exome v3 n.d.
<http://sequencing.roche.com/products/nimblegen-seqcap-target-enrichment/seqcap-ez-system/seqcap-ez-exome-v3.html> (accessed July 24, 2017).

- 90 Agilent SureDesign n.d. <https://earray.chem.agilent.com/suredesign/> (accessed July 24, 2017).

- 91 Kieleczawa J. Fundamentals of sequencing of difficult templates--an overview. *J Biomol Tech* 2006;17:207–17.

- 92 Katoh M, Katoh M. Identification and characterization of human HES2, HES3, and HES5 genes in silico. *Int J Oncol* 2004.
doi:10.3892/ijo.25.2.529.

- 93 Kita A, Imayoshi I, Hojo M, Kitagawa M, Kokubu H, Ohsawa R, et al. *Hes1* and *Hes5* Control the Progenitor Pool, Intermediate Lobe Specification, and Posterior Lobe Formation in the Pituitary Development. *Mol Endocrinol* 2007;21:1458–66. doi:10.1210/me.2007-0039.

- 94 Rizzoti K, Brunelli S, Carmignac D, Thomas PQ, Robinson IC, Lovell-Badge R. SOX3 is required during the formation of the hypothalamo-pituitary axis. *Nat Genet* 2004;36:247–55. doi:10.1038/ng1309.

- 95 Pérez Millán MI, Vishnopoliska SA, Daly AZ, Bustamante JP, Seilicovich A, Bergadá I, et al. Next generation sequencing panel based on single

- molecule molecular inversion probes for detecting genetic variants in children with hypopituitarism. *Mol Genet Genomic Med* 2018;6:514. doi:10.1002/mgg3.395.
- 96 Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol* 2011;12:R95. doi:10.1186/gb-2011-12-9-r95.
- 97 Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. *Genome Biol* 2011;12:R97. doi:10.1186/gb-2011-12-9-r97.

Anexos

Anexos

Anexo 1. Tabelas com média da nota Phred para cada amostra nas quatro tecnologias.

Tabela 1 e 2. Notas Phred para Nextera e SureSelect.

Amostra	Média Phred
1210	33,6
1211	33,4
1212	33,4
1629	34,7
1985	34,8
2053	34,7
2078	34,7
2260	34,9
2261	34,9
2262	34,9
10654	34,4
10655	34,2
10656	34,5
10867	34,4
10868	34,4
11004	34,7
11005	34,7
11088	34,7
116	34,9
2298	35,4

Tabela 3 e 4. Notas Phred SureSelect+UTR e

Amostra	Média Phred
7709	35,4
7846	35,4
9378	35,5
10266	35,4
10317	35,4
10460	35,4
6950	34,2
6951	34,2
6952	34,2
1493	34,2
1494	34,2
3028	34,2
3029	34,2
334N	34,6
335	34,5
336	34,4
10652	34,5
10658	34,5
3009	34,6
NimbleGen 3010	34,4

Anexo 2. Tabelas com nome dos genes com 100% de sua extensão coberta em 1x, 10x e 20x, respectivamente.

Tabela 1. Genes com 100% de cobertura em sua extensão com 1x de reads.

Nextera						
HNRNPU	WNT4	GPR161	SIX3	HESX1	SOX2	POU1F1
CTNNB1	PITX2	RBPJ	ISL1	FGF10	PROP1	RBM28
GHRHR	GLI3	FGFR1	WDR11	BMPR1A	POLR3A	GATA3
PAX6	CDON	BMP4	SIX6	OTX2	ARNT2	GH1
CDH2	TGIF1	TCF4	AES	PNPLA6	PROKR2	GHRH
BMP7	NOTCH2					
NimbleGen						
NOTCH2	LHX4	HNRNPU	GPR161	RNPC3	IFT172	TCF7L1
HES1	GHSR	HESX1	SOX2	GATA2	POU1F1	CTNNB1
SureSelect						
NOTCH2	LHX4	HNRNPU	GPR161	IFT172	SIX3	HES1
SOX2	GATA2	POU1F1	CTNNB1	WNT5A	HHIP	PITX2
RBPJ	ISL1	FGF10	FGF18	PROP1	SOX4	RBM28

SHH	GHRHR	GLI3	FGFR1	CHD7	LHX3	LHX2
TLE1	HDAC6	SOX3	FGF8	BMPR1A	POL3A	GATA3
PAX6	CDON	TBX3	BMP4	SIX6	OTX2	DMXL2
ARNT2	TLE3	TBX2	GH1	CDH2	TGIF1	RAX
TCF4	PNPLA6	PROKR2	GHRH	BMP2	BMP7	
SureSelect+UTR						
NOTCH2	LHX4	HNRNPU	GPR161	IFT172	SIX3	HES1
SOX2	GATA2	POU1F1	CTNNB1	WNT5A	HHIP	PITX2
RBPJ	ISL1	FGF10	FGF18	PROP1	SOX4	RBM28
SHH	GHRHR	GLI3	FGFR1	CHD7	LHX3	LHX2
HDAC6	SOX3	BMPR1A	POL3RA	GATA3	PAX6	CDON
TBX3	BMP4	SIX6	OTX2	DMXL2	ARNT2	TLE3
TBX2	GH1	CDH2	TGIF1	RAX	TCF4	PROKR2
GHRH	BMP2	BMP7				

Tabela 2 e 3. Genes com 100% de cobertura em sua extensão com 10x e 20x de reads.

Nextera					
POU1F1	ISL1	GHRHR	PAX6	GH1	TCF4
NimbleGen					
GPR161	RNPC3	HES1	HESX1	POU1F1	CTNNB1
SureSelect					
SOX2	LHX4	FGF10	GPR161	WNT5A	FGF18
GHRHR	GATA3	PAX6	CTNNB1	BMP4	SIX6
TGIF1	TCF4	GHRH	BMP2	HES1	ARNT2
SureSelect+UTR					
SOX2	FGF18	PAX6	GPR161	IFT172	SIX3
BMP2	HES1				
NimbleGen					
HES1					
SureSelect					
GPR161	FGF18	GHRH			

Anexo 3. Genes com piores coberturas em cada tecnologia nos cortes de 1x, 10x e 20x, respectivamente.

Tabela 1. Genes com piores coberturas considerando 1x de reads.

Gene	Cobertura 1x (%)	Cobertura 10x (%)	Cobertura 20x (%)	Tecnologia
HES5	85,6	68	54,4	Nextera
HES5	91,9	39,5	23,3	NimbleGen
HES5	97,6	88,2	82,8	SureSelect
RAX	90,8	42,3	26,7	NimbleGen
AES	98,3	90,3	94,9	SureSelect

AES	96,8	82,4	64,5	SureSelect+UTR
SOX3	92,1	61,1	43,4	NimbleGen
SOX4	93,1	67,5	58	Nextera
FGF8	93,4	84,8	79,4	NimbleGen
FGF8	99,3	87,3	72,1	SureSelect+UTR
TBX2	93,4	73,9	56,5	Nextera
ZSWIM6	93,5	79,7	78,6	NimbleGen
ZSWIM6	95	87,7	81,4	Nextera
ZSWIM6	98,8	96,1	93,3	SureSelect
ZSWIM6	98,4	93,1	85,5	SureSelect+UTR
WNT5A	96,9	81,9	73,2	Nextera
KAL1	99,6	91	85,5	SureSelect
KAL1	98,9	88,7	80,2	SureSelect+UTR
WNT4	99,7	95,4	93,1	SureSelect
WNT4	98,3	92,7	90,8	SureSelect+UTR

Tabela 2. Genes com piores coberturas considerando 10x de reads.

Gene	Cobertura 1x (%)	Cobertura 10x (%)	Cobertura 20x (%)	Tecnologia
SOX4	93,1	67,5	58	Nextera
HES5	85,6	68	54,4	Nextera
HES5	91,9	39,5	23,3	NimbleGen
HES5	97,6	88,2	82,8	SureSelect
HES5	99,8	82	61,8	SureSelect+UTR
TBX2	93,4	73,9	56,5	Nextera
SOX3	97,6	74,4	56,2	Nextera
SOX3	92,1	61,1	43,4	NimbleGen
WNT5A	98,9	81,9	73,2	Nextera
RAX	90,8	42,3	26,7	NimbleGen
LHX3	93,8	61,1	36,2	NimbleGen
SHH	97,1	71,9	50,3	NimbleGen
AES	98,3	90,3	94,9	SureSelect
AES	96,8	82,4	64,5	SureSelect+UTR
KAL1	99,6	91	85,5	SureSelect
KAL1	98,9	88,7	80,2	SureSelect+UTR
RNPC3	99,4	92	76,2	SureSelect
RNPC3	99,9	86,3	59,1	SureSelect+UTR
FGF8	99,3	87,3	72,1	SureSelect+UTR
FGF8	100	92,7	94,6	SureSelect

Tabela 3. Genes com piores coberturas considerando 20x de reads.

Gene	Cobertura 1x (%)	Cobertura 10x (%)	Cobertura 20x (%)	Tecnologia
HES5	85,6	68	54,4	Nextera

HES5	91,9	39,5	23,3	NimbleGen
HES5	97,6	88,2	82,8	SureSelect
HES5	99,8	82	61,8	SureSelect+UTR
SOX3	97,6	74,4	56,2	Nextera
SOX3	92,1	61,1	43,3	NimbleGen
TBX2	93,4	73,9	56,5	Nextera
SOX4	93,1	67,5	58	Nextera
WNT5A	96,9	81,9	73,2	Nextera
RAX	90,8	42,3	26,7	NimbleGen
RAX	100	91,5	72	SureSelect+UTR
LHX3	93,8	61,1	36,2	NimbleGen
SHH	97,1	71,9	50,3	NimbleGen
RNPC3	99,9	92	76,2	SureSelect
RNPC3	99,9	86,3	59,1	SureSelect+UTR
FGF8	100	92,7	84,6	SureSelect
FGF8	99,3	87,3	72,1	SureSelect+UTR
AES	98,3	90,3	84,9	SureSelect
AES	96,8	82,4	64,5	SureSelect+UTR
KAL1	99,6	91	85,5	SureSelect