UNIVERSIDADE DE SÃO PAULO
FFCLRP - DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

Modelagem e informatização do processo de análise do sequenciamento de exoma humano

Marcelo Gomes de Paula

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP, como parte das exigências para a obtenção do título de Mestre em Ciências, Área: Computação Aplicada

Ribeirão Preto - SP

2018

MARCELO GOMES DE PAULA

Modelagem e informatização do processo de análise do sequenciamento de exoma humano

Versão Revisada

(Versão original encontra-se na unidade que
aloja o Programa de Pós-graduação)

Dissertação apresentada à Faculdade de
Filosofia, Ciências e Letras de Ribeirão Preto
da USP, como parte das exigências para a
obtenção do título de Mestre em Ciências,
Área: Computação Aplicada

Orientador: Prof. Dr. Joaquim Cezar Felipe

Ribeirão Preto - SP

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Paula, Marcelo Gomes de

Modelagem e informatização do processo de análise do sequenciamento de exoma humano. Ribeirão Preto, 2018.

100 p. : il. ; 30 cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Neurologia.

Orientador: Felipe, Joaquim Cezar

1. Exoma. 2. Ferramentas bioinformática. 3. Modelagem conhecimento

Marcelo Gomes de Paula

Modelagem e informatização do processo de análise do sequenciamento de exoma humano

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Mestre em Ciências.

Aprovado em: 10 de outubro de 2018

Banca Examinadora

Prof. Dr.	<u>Ricardo Rodrigues Ciferri</u>
Instituição:	<u>Universidade Federal de São Carlos, Centro de Ciências Exatas e de Tecnologia</u>
Julgamento:	<u></u>

Prof. Dr.	<u>Ricardo Zorzetto Nicoliello Vêncio</u>
Instituição:	<u>Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras</u>
Julgamento:	<u></u>

Prof. Dr.	<u>Wilson Araújo da Silva Junior</u>
Instituição:	<u>Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto</u>
Julgamento:	<u></u>

DEDICATÓRIA

A minha noiva e a meus familiares, com amor, admiração e gratidão por sua compreensão, carinho, presença e incansável apoio ao longo do período de elaboração deste trabalho.

AGRADECIMENTOS

Ao Prof. Dr Joaquim Cezar Felipe pela orientação, incentivo e compreensão.

Ao pessoal do Laboratório de Genética Molecular e Bioinformática pelo apoio e incentivo durante esses anos. Em especial Raul, Jessica, Prof. Wilson, Danielle, Simone e Josane.

Aos meus pais, Moacir e Eliane, e aos meus irmãos Luciana e Leandro e a minha noiva Gisele pelo inestimável apoio durante minha vida acadêmica e sempre.

Aos meus colegas e amigos que tiveram paciência comigo em momentos de tensão e de empenho.

Aos meus amigos de mestrado por compartilhar experiências e conhecimentos.

Por fim, agradeço aos professores e colaboradores do Programa de Pós-Graduação de Computação Aplicada, do Departamento de Computação e Matemática da FFCLRP-USP.

RESUMO

O sequenciamento de nova geração tornou-se recentemente uma opção economicamente viável para a realização de aconselhamento genético, tendo proporcionado oportunidades sem precedentes para a pesquisa genômica personalizada, uma vez que o sequenciamento permite identificar a presença de variantes patogênicas. O teste genético que permite identificar as variantes patogênicas possui procedimentos laboratoriais como a coleta, preparação e sequenciamento do material e outros computacionais como o alinhamento, anotação e análise do exoma. Apesar do fato de que a grande quantidade de dados gerados no sequenciamento torna a atividade de análise relativamente complexa, esta tem sido realizada de forma manual por pesquisadores e profissionais da saúde, apesar de existir uma variedade de ferramentas computacionais disponíveis para auxiliar a análise e a interpretação de variações genéticas pontuais. Isto acontece porque a maioria destas ferramentas são generalistas, pouco amigáveis e não atendem às necessidades em pesquisa e atendimento clínico, não sendo utilizadas rotineiramente nesses ambientes. A análise manual, por sua vez, caracteriza-se por gerar resultados altamente dependentes tanto do conhecimento do pesquisador quanto da especificidade do procedimento realizado. O presente trabalho tem por objetivo desenvolver a modelagem do processo de análise de exoma, considerando a presença e a interação de diferentes papéis de usuário, num processo padronizado, porém flexível, juntamente com a implementação de uma ferramenta computacional amigável para a execução da análise baseada na modelagem proposta de forma organizada e otimizada, oferecendo suporte a atividades relacionadas à pesquisa e ao atendimento clínico. As análises processadas com a ferramenta também têm seus parâmetros e resultados armazenados e disponibilizados para uso em investigações futuras, possibilitando o compartilhamento do conhecimento para geração de novas descobertas. Nas avaliações realizadas para validação da proposta o modelo e a ferramenta mostram-se eficientes e apresentam boa usabilidade para a realização das análises.

Palavras-chave: Exoma. Ferramentas bioinformática. Modelagem conhecimento.

ABSTRACT

New generation sequencing has recently become an economically viable option for genetic counseling and has provided unprecedented opportunities for personalized genomic research since sequencing allows the identification of the presence of pathogenic variants. The genetic test that allows to identify the pathogenic variants has laboratory procedures like the collection, preparation and sequencing of the material and other computational as the alignment, annotation and analysis of the exoma. In spite of the fact that the large amount of data generated in the sequencing makes the analysis activity relatively complex, it has been performed manually by researchers and health professionals, although a variety of computational tools are available to aid analysis and the interpretation of specific genetic variations. This is because most of these tools are generalist, unfriendly and do not meet clinical research and care needs and are not routinely used in these settings. The manual analysis, in turn, is characterized by generating results highly dependent on both the knowledge of the researcher and the specificity of the procedure performed. The present work aims to develop the modeling of the exoma analysis process, considering the presence and interaction of different user roles, in a standardized but flexible process, together with the implementation of a user-friendly computational tool for performing the analysis based in the proposed modeling in an organized and optimized way, supporting activities related to research and clinical care. The analyzes processed with the tool also have their parameters and results stored and made available for use in future investigations, allowing the sharing of knowledge to generate new discoveries. In the evaluations carried out to validate the proposal, the model, and the tool are efficient and have good usability for carrying out the analysis.

Keywords: Exoma. Bioinformatics tools. Knowledge modeling.

LISTA DE FIGURAS

Figura 1 - Teste genético	21
Figura 2 - Introns e Exons	23
Figura 3 - Exemplo de arquivo VCF	30
Figura 4 - Exemplo de arquivo SAM	31
Figura 5 - Arquivo Fasta	32
Figura 6 - Exemplo de arquivo fastq	33
Figura 7 - A arquitetura e as funcionalidades do ClinLabGeneticist	37
Figura 8 - fluxo de trabalho do VariantStudio	38
Figura 9 - Illumina VariantStudio	39
Figura 10 - Visualização do pacientes no Mendel,MD	41
Figura 11 - Primeira etapa na seleção de variantes do Mendel,MD	41
Figura 12 – Fases do modelo de análise de NGS	45
Figura 13 - Filtro padrão	47
Figura 14 - Fluxo de trabalho do NGA	53
Figura 15 – Atividades do gestor e revisor na NGA	54
Figura 16 – Diagrama de caso de uso de NGA	55
Figura 17 - Login no NGA	56
Figura 18 - Dashboard do administrador.....	56
Figura 19 - Cadastrado de usuário	57
Figura 20 - Gerenciar usuários	58
Figura 21 - Alerta de erro	59
Figura 22 - Cadastro de população.....	60
Figura 23 - Cadastro de banco de dados biológico	61
Figura 24 - Manutenção de genes	62
Figura 25 - Manutenção das doenças	63
Figura 26 - Manutenção das informações biológicas	64
Figura 27 – Importar informações biológicas	65
Figura 28 - Opções	66
Figura 29 - Cadastro de paciente	67
Figura 30 – Importação do arquivo de variantes	68
Figura 31 - Manutenção dos filtros	72
Figura 32 – Manutenção dos filtros, com gene e posição	73
Figura 33 - Nova análise	74
Figura 34 - Seleção de variantes	75
Figura 35 - Modificar opções do filtro.....	76
Figura 36 - Filtros aplicados	77
Figura 37 - Seleção de revisores	78
Figura 38 - Status da análise	78
Figura 39 – Finalizar laudo da análise	79
Figura 40 - Informações apoiar o laudo	80
Figura 41 - Exemplo de laudo	80
Figura 42 - Visualização das variantes pelo revisor	81
Figura 43 - Classificação da variante pelo gestor	82

Figura 44 - Camadas físicas do NGA	84
Figura 45 - Camadas lógicas do NGA	85
Figura 46 – Criação da análise de validação.....	90
Figura 47 - Seleção das variantes na análise de validação	91

LISTA DE TABELAS

Tabela 1 - Campos do arquivo SAM	30
Tabela 2 - Descrição do arquivo de população	60
Tabela 3 – Descrição do arquivo de genes	62
Tabela 4 – Descrição do arquivo de doenças	63
Tabela 5 - Descrição do arquivo de informações biológicas	65
Tabela 6 - Links de apoio ao revisor	82
Tabela 7 - Questões aplicadas na avaliação de usabilidade.....	87
Tabela 8 - Respostas de cada usuário para o questionário de usabilidade	88

LISTA DE QUADROS

Quadro 1 - Comparativo entre os trabalhos correlatos e a NGA.....	43
Quadro 2 - Filtros disponíveis.....	49
Quadro 3 - Descrição do arquivo de variantes	68

LISTA DE ABREVIATURAS E SIGLAS

BAM	<i>Binary Alignment Map</i> (Mapa de Alinhamento Binário)
BWA	<i>Burrows-Wheeler Aligner</i> (Alinhador de Burrows-Wheeler)
CDNA	<i>complementary DNA</i> (DNA complementar)
CMG	Centro de Medicina Genômica
COSMIC	<i>Catalogue of somatic mutations in cancer</i> (Catálogo de mutações somáticas de câncer)
CSV	<i>Comma-separated values</i> (Valores separados por vírgula)
DAO	<i>Data access object</i> (Objeto de acesso a dados)
DNA	<i>Deoxyribonucleic acid</i> (Ácido desoxirribonucleico)
GATK	<i>Genome Analysis Tool Kit</i> (Kit de ferramentas de análise de genoma)
HCFMRP	Hospital das clínicas da faculdade de medicina de Ribeirão Preto - USP
HPO	<i>Human Phenotype Ontology</i> (Ontologia do Fenótipo Humano)
HUGO	<i>Human Genome Organisation</i> (Organização do Genoma Humano)
ICD	<i>International Classification of Diseases</i> (Classificação Internacional de Doenças)
JPA	<i>Java Persistence API</i> (API de persistência Java)
JSF	<i>JavaServer Faces</i>
MRNA	<i>Messenger RNA</i> (RNA de mensageiro)
MVC	<i>Model view controller</i> (modelo visão controlador)
NA	Não se aplica
NCBI	<i>National Center for Biotechnology Information</i> (Centro Nacional de Informação Biotecnológica)
NCRNA	non-coding RNA (RNA não-codificante)
NGA	Next-Gen Analyser (Analisador de nova geração)
NGS	<i>Next Generation Sequencing</i> (Sequenciamento de nova geração)

OMIM	<i>Online Mendelian Inheritance in Man</i> (Herança Mendeliana no Homem Online)
PCR	<i>Polymerase Chain Reaction</i> (Reação em cadeia da polimerase)
REST	<i>Representational State Transfer</i> (Transferência de Estado Representacional)
SAM	<i>Sequence Alignment Map</i> (Alinhamento de sequência/Mapa)
SGBD	Sistema de Gerenciamento de Banco de Dados
SIFT	<i>Scale invariant feature transform</i>
SNP	<i>Single nucleotide polymorphism</i> (Polimorfismo de nucleotídeo único)
SP	<i>Stored procedure</i> (Procedimento armazenado)
UCSC	<i>University of California, Santa Cruz</i> (<i>Universidade da Califórnia em Santa Cruz</i>)
VCF	<i>Variant Call Format</i> (Formato de Chamada Variante)
VEP	<i>Variant Effect Predictor</i> (preditor do efeito da variante)

SUMÁRIO

1 INTRODUÇÃO	16
1.1 Motivação e objetivos	18
1.2 Objetivos específicos.....	19
1.3 Organização da dissertação	20
2 FUNDAMENTOS TEÓRICOS.....	21
2.1 Sequenciamento de DNA e testes genéticos	22
2.2 <i>Pipeline</i> de Pré-processamento	25
2.3 Banco de dados biológicos.....	25
2.3.1 Omim	26
2.3.2 Orphanet	26
2.3.3 HPO	27
2.3.4 Hugo.....	27
2.3.5 1000 Genomas	28
2.3.6 ICD.....	28
2.4 Formatos de arquivos.....	28
2.4.1 VCF.....	28
2.4.2 SAM / BAM.....	30
2.4.3 Fasta.....	31
2.4.4 Fastq.....	32
2.5 Ferramentas para alinhamento e anotação.....	33
2.5.1 BWA.....	33
2.5.2 Picard.....	34
2.5.3 GATK	34
2.5.4 snpEFF.....	34
2.5.5 VEP	34
2.5.6 UMD Predictor	35
3 TRABALHOS CORRELATOS	36
3.1 ClinLabGeneticist.....	36
3.2 VariantStudio.....	38

3.3 Mendel,MD.....	40
3.4 Comparativo	42
4 PROPOSTA DO TRABALHO.....	45
4.1 Anotação.....	46
4.2 Seleção das variantes.....	46
4.2.1 Filtro padrão	46
4.2.2 Filtros.....	47
4.4 Revisão das variantes	52
4.5 Laudo	52
5 NEXT-GEN ANALYSER	53
5.1 Administrador.....	56
5.1.1 Gerenciamento dos usuários.....	56
5.1.2 Gerenciamento das populações	59
5.1.2 Gerenciamento do banco de dados	60
5.1.3 Gerenciamento dos genes.....	61
5.1.4 Gerenciamento das doenças.....	63
5.1.5 Gerenciamento das informações biológicas	64
5.1.6 Opções.....	65
5.2 Gestor.....	66
5.2.1 Cadastro de paciente	66
5.2.2 Filtro	71
5.2.3 Análise	73
5.2.4 Seleção de variantes.....	74
5.2.5 Laudo da análise	79
5.3 Revisor	81
5.3.1 Classificação das variantes.	81
5.4 Arquitetura	83
5.5 Implementação do sistema.....	83
6 VALIDAÇÃO	86
6.1 Avaliação da usabilidade	86
6.2 Estudo de caso.....	90
6.3 Considerações finais	92
7 CONCLUSÃO.....	93

Trabalhos futuros	94
--------------------------------	-----------

1 INTRODUÇÃO

As técnicas de sequenciamento de nova geração (*Next Generation Sequencing* - NGS) tornaram o sequenciamento de genomas mais rápido e econômico, permitindo examinar milhares de genomas humanos e catalogar as variações nas sequências de nucleotídeos que aumentam o risco de várias doenças, como, por exemplo, o câncer e o autismo (JOHNSON e colab., 2017).

A identificação de uma sequência causadora de uma doença permite o correto direcionamento em casos de tratamento. O sequenciamento do exoma, na maioria das vezes, detecta uma quantidade alta de sequências por amostra, exigindo uma análise mais aprofundada para identificar as poucas variantes associadas a uma determinada doença. Esta é comumente chamada de análise de exoma.

Diferentes abordagens podem ser usadas de acordo com as necessidades e as questões a serem avaliadas nos testes genéticos. Um possível fluxo de teste genético consiste nos seguintes processos: processamento da amostras, sequenciamento, análise da qualidade, alinhamento, realização da chamada de variante, anotação, filtragem e emissão do laudo (KAMPS e colab., 2017).

No processamento da amostra é realizado a coleta e preparação da amostra que será sequenciada. O sequenciamento, ou seja, a identificação da sequência de nucleotídeos, é realizado e estas sequências são armazenadas em um ou mais arquivos do tipo fastq.

O controle geral sobre a qualidade de leitura é realizado para remoção de leituras de má qualidade e, na etapa seguinte, as sequências de nucleotídeos são alinhadas contra uma referência, onde apenas as sequências que possuem alguma mutação, ou seja, alguma divergência em relação à referência, são selecionadas no processo. E estas sequências que possuem mutações são chamadas de variantes.

As variantes são então anotadas, isto é se tenta identificar seu gene e compreender suas funções e impacto (JOHNSON e colab., 2017). A última etapa deste processo consiste em selecionar dentre todas as variantes anotadas quais são as causadoras da doença em estudo, para posterior emissão do laudo do teste genético. Este procedimento, inerentemente complexo e dependente do especialista consiste no que está sendo chamado de análise do exoma.

Em geral um *pipeline* é construído para realizar o processo de verificação da qualidade, alinhamento e anotação das variantes, utilizando ferramentas já disponíveis para esta atividade. Porém, o processo de análise, responsável pela seleção das mutações, é realizado de forma

manual, fazendo com que interpretar e gerenciar o grande conjunto de dados de resultados produzido pelos NGS continua sendo um desafio (WANG e colab., 2015).

Para realizar a análise de forma manual, os pesquisadores e/ou profissionais da saúde fazem uso de uma planilha eletrônica, exportando um arquivo no formato *Variant Call Format* (VCF) para ela, sem padronização do processo, ficando a cargo dele decidir quais critérios utilizar, estando muitas vezes limitado pelo seu conhecimento em manipulação de planilha eletrônica e pela sua experiência em relação ao processo. Em geral o especialista deve consultar base de dados públicas que contém informações sobre o assunto, como por exemplo *Mendelian Inheritance in Man* (OMIM), Orphanet e *Human Phenotype Ontology* (HPO).

Por outro lado, existem ferramentas computacionais disponíveis para análise, interpretação e seleção das mutações tais como ClinLabGeneticist, VariantStudio e Mendel.MD. Em geral estas ferramentas permitem ao usuário importar um arquivo no formato VCF, que contém as variantes identificadas a partir do sequenciamento de genomas e filtrar estas informações. Todavia, a maioria destas ferramentas são projetadas principalmente para fins de pesquisa e não atendem todas as necessidades de um laboratório que trabalha com pesquisa e atendimento clínico (WANG et al., 2015). Tanto os laboratórios de pesquisa quanto aqueles de atendimento clínico necessitam que o gerenciamento de dados e o fluxo do trabalho seja padronizado e automatizado, permitindo armazenamento, classificação e compartilhamento de variantes, acessível por uma interface gráfica intuitiva ao usuário.

Neste trabalho é descrito o desenvolvimento da modelagem do processo de análise de exoma, que compreende em realizar mais anotações, o processo de seleção e laudo das variantes. A inclusão de mais passos no processo de anotação realizado pela modelagem proposta é para permitir, por exemplo, a seleção automática de genes por doenças baseado em banco de dados públicos, que podem ser importados e acessados pela ferramenta como por exemplo o OMIM, o Orphanet e o HPO. A modelagem foi construída a partir de uma série de reuniões realizadas com especialistas para levantamento de requisitos e aquisição de conhecimento sobre o processo. O objetivo é padronizar o processo, de forma a torná-lo menos dependente do usuário, porém oferecendo a este usuário a possibilidade de construir análises personalizadas.

O sistema computacional chamado de Next-Gen Analyser (NGA) foi implementado seguindo a modelagem do processo.

Para os procedimentos que precedem a análise realizada pelo NGA, um *pipeline* já estabelecido é utilizado.

Este *pipeline*, que chamamos de *pipeline* de pré-processamento, produz como resultado

um arquivo com as variantes já anotadas. Este arquivo é que é utilizado como entrada para a ferramenta NGA construída neste trabalho.

Os usuários do NGA podem pertencer a um ou mais perfis de usuário. Os perfis disponíveis são: administrador, gestor e revisor.

O administrador é quem gerencia os usuários e as bases de dados do sistema.

O gestor é o responsável pela análise do exoma, é a pessoa que irá criar uma análise e selecionar as variantes suspeitas. Para realizar esta atividade, o gestor conta com um banco de filtros padrão que aplica os critérios definidos pelo especialista; porém, caso desejado, o gestor poderá modificar estes filtros ou mesmo criar seus próprios filtros.

O revisor é quem irá apoiar o gestor na análise, ele receberá a lista de variantes do gestor e cada variante destas listas deve ser classificada de acordo com a patogenicidade, o revisor conta com *links* para os bancos de dados cadastrados pelo administrador para consultar de forma simples as informações *online* destas variantes.

O laudo é emitido pelo gestor, que utiliza as informações fornecidas pelos revisores designados por ele para realizar esta atividade.

Todos estes passos são organizados em um fluxo de trabalho que facilita a comunicação entre o gestor e o revisor.

Os dados resultantes das análises realizadas com a NGA são armazenados de maneira padronizada, permitindo que no futuro tais dados sejam utilizados em novas investigações em buscas por novos conhecimentos. Como exemplo, podemos citar a descoberta de associação de uma mutação específica a uma determinada patologia.

1.1 Motivação e objetivos

Doenças causadas por mutação no DNA (ácido desoxirribonucleico) podem ser transmitidas de uma geração a outra. A anemia falciforme é um exemplo de doença grave causada pela mutação de um único gene (REES e colab., 2010).

A tecnologia de NGS permite a leitura de milhões nucleotídeos e o nucleotídeo complementar, chamados de pares de bases, em uma única corrida, apresentando custo e tempo menor que a tecnologia de Sanger (CARVALHO e SILVA, 2010). NGS, além de ser utilizado em pesquisas genômicas, também pode ser utilizado como ferramenta no diagnóstico molecular de grande parte das doenças genéticas.

A interpretação dos dados de NGS não é uma tarefa trivial, e continua sendo um desafio devido às dificuldades ligadas principalmente ao grande volume de dados. Quando se deseja

identificar se uma variante está associada a uma determinada patologia, existem bancos de dados que contém esta informação, porém é necessário para cada variante pesquisar em um destes bancos, como por exemplo o OMIM. Esta atividade acaba demandando muito tempo e está propensa a gerar erros de marcação quando realizada manualmente.

Este trabalho tem por objetivo desenvolver a modelagem do processo de análise de exoma, assim como a implementação de uma ferramenta computacional para a execução dela. O sistema funcionará não só como um fluxo de trabalho, guiando os passos da análise, mas também automatizando parte das atividades relacionadas à localização de variantes associadas a uma determinada patologia.

Com o uso do sistema, erros de marcação humana podem ser eliminados, consultas a bases de dados e classificação das variantes a partir do conhecimento extraído de bases de dados públicas tornam-se simples e eficientes, por permitir uma análise com um procedimento guiado, reduzindo a dependência do pesquisador na interpretação dos dados.

1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

- I. Modelar o processo de análise de variantes presentes no exoma humano.

A modelagem do processo de análise foi realizada a partir do levantamento de requisitos realizado em reuniões com especialistas, que definiram as principais etapas do *pipeline*, os filtros considerados na análise, assim como as possíveis variações de análise que podem ser desejadas pelos usuários.

- II. Desenvolver a ferramenta NGA, que implementa o modelo proposto.

O desenvolvimento da ferramenta foi realizado utilizando a linguagem Java e o framework Java Server Faces, seguindo a metodologia Scrum, alcançando maior qualidade do produto e transparência no processo do projeto.

- III. Realizar a avaliação da NGA.

A validação da NGA foi realizada efetuando a análise em experimentos já analisados previamente de forma manual, cujo resultados foram confirmados experimentalmente. O resultado obtido com a análise realizada na ferramenta proposta foi comparado aos resultados conhecidos e validados. Um questionário de usabilidade foi respondido por um conjunto de usuários especialistas, para indicar o nível de satisfação no uso da ferramenta.

- IV. Disponibilizar a NGA para a comunidade científica.

1.3 Organização da dissertação

A presente dissertação está organizada da seguinte forma: No Capítulo 2 são descritos os fundamentos teóricos que serviram de base para o desenvolvimento deste trabalho. No Capítulo 3 é apresentado os trabalhos correlatos: ClinLabGeneticist, VariantStudio e Mendel,MD. No capítulo 4 o modelo proposto no trabalho é apresentado. No capítulo 5 é descrito a ferramenta NGA que implementa o modelo proposto e o processo de validação da ferramenta. Por fim, no Capítulo 6 a conclusão do trabalho, com a discussão dos resultados e trabalhos futuros e as considerações finais.

2 FUNDAMENTOS TEÓRICOS

O NGS produz um grande volume de dados, disponibilizando a pesquisadores e profissionais da saúde um conjunto de milhares de variantes genéticas. A análise de NGS deve selecionar um conjunto de variantes as quais possam ser potencialmente prejudiciais à saúde do paciente e que possam contribuir para os pesquisadores e profissionais da saúde no diagnóstico clínico.

No teste genético, após a amostra ser sequenciada é verificada a qualidade da sequência, e a mesma é alinhada contra uma referência, gerando as variantes, as quais então são anotadas para que a análise do exoma seja realizada. Na análise do exoma devem ser selecionadas as variantes patogênicas para que o laudo possa ser emitido (KAMPS e colab., 2017). Na *Figura 1* é possível visualizar este fluxo do teste genético.

Figura 1 - Teste genético

Dentro deste contexto, este trabalho visou a modelado da análise do exoma (a última etapa do processo) e o desenvolvimento da ferramenta Next-Gen Analysis (NGA), cujos

objetivo é apoiar os pesquisadores e profissionais da saúde nas atividades que compreendem parte do processo de anotação, juntamente com a filtragem das variantes e a emissão do laudo. A NGA utiliza como entrada as variantes que já estejam alinhadas e anotadas, ou seja, onde os processos anteriores do teste genéticos já devem ter sido realizados. Para atender a este requisito, os dados sequenciados são previamente submetidos a um *pipeline* já estabelecido que realiza todo o pré-processamento computacional anterior à análise.

O *pipeline* de pré-processamento realiza os procedimentos de avaliação da qualidade, alinhamento, chamada de variantes e anotações, chamando ferramentas como BWA e snpEFF.

Esta seção apresenta resumidamente os fundamentos teóricos sobre os temas abordados neste trabalho, além apresentar as ferramentas utilizada no *pipeline* de pré-processamento.

2.1 Sequenciamento de DNA e testes genéticos

Os seres humanos são compostos de células diploides que contém pares de cromossomo (22 pares, mais o XX ou XY). Os cromossomos são estruturas que armazenam de forma compactada o ácido desoxirribonucleico (DNA), sendo o DNA um polímero linear composto por uma fita dupla de nucleotídeos, formado por quatro bases nitrogenadas, que são: adenina (A), citosina (C), timina (T), guanina (G). Um lado da fita do DNA não possui a hidroxila do carbono-5 e do outro lado não possui a hidroxila do carbono-3, por isso os lados da linha são identificados respectivamente por 5' e 3'. As diferentes combinações das quatro bases nitrogenadas é a base que produz a variabilidade dos seres vivos. O DNA é transcrito em uma molécula molde para a formação das proteínas, denominada ácido ribonucleico (RNA), desta forma, podendo agir de forma livre no organismo. Assim a região que possui um potencial de transcrever um RNA é chamada de gene.

O gene é composto por éxons e íntrons. Os éxons são regiões codificadora de proteínas, já os íntrons são as regiões de um gene que não codificam qualquer parte da proteína produzida pelo gene, a junção de todos os éxons é o que chamamos de RNA mensageiro (mRNA) que será usado para a produção das proteínas (*Figura 2*).

Figura 2 - Introns e Exons

Erros na replicação do DNA e transcrição do DNA para RNA são relativamente comuns e caso ocorram nas células reprodutoras de um organismo, eles podem ser transmitidos aos seus descendentes. Estas alterações na sequência de DNA são chamadas de mutações, que podem ser ou não prejudiciais ao desenvolvimento e sobrevivência de um organismo.

Sequenciamento de DNA é um processo utilizado para se determinar a ordem linear dos nucleotídeos em um fragmento de DNA, que pode ser feito por diversas técnicas. O sequenciamento do DNA pode ajudar a identificar genes associados a algumas patologias e verificar a presença de mutação em um indivíduo cuja patologia já tem uma mutação conhecida. Este resultado do sequenciamento do DNA apoia um aconselhamento genético. Para o modelo proposto qualquer um dos métodos de sequenciamento como por exemplo Illumina ou Íon Torrent pode ser aplicado, desde que a saída deste processo seja o arquivo fastq contendo a descrição dos fragmentos de DNA do paciente.

Localizar e identificar genes específicos é o primeiro passo para revelar as mutações que alteram a informação genética e que podem causar doenças. A partir destas informações é possível realizar testes para diagnósticos, bem como levar a novas formas de tratamento e eventualmente a cura completa da anomalia genética (FARAH, 2007).

Atualmente, os testes genéticos podem ser realizados em: diagnóstico pré-natal, detecção de indivíduos portadores do gene, que podem ser em casos assintomáticos visando o

aconselhamento genético ou ainda em programas de triagem populacional (FARAH, 2007). O diagnóstico pré-natal permite determinar se o feto está ou não sendo afetado por alguma doença genética conhecida.

Hipertensão, diabetes e alguns tipos de câncer são exemplos de patologia determinadas por herança multifatorial, em que o fenótipo é resultado da interação entre vários genes com diversos fatores ambientais, neste caso, a presença de genes que aumentam a predisposição genética. A predisposição genética não significa que a ocorrência da doença seja inevitável. Ao contrário, esse indivíduo pode reconsiderar, se necessário, seus hábitos de estilo de vida de forma a prevenir o aparecimento dos sintomas. Por exemplo, uma pessoa com propensão genética à hipertensão poderia praticar exercícios físicos, ter uma dieta mais saudável e ter um acompanhamento médico mais cuidados para evitar problemas futuros. Nesse caso, mesmo se não existir cura disponível para uma doença de herança multifatorial, o teste genético auxilia na prevenção.

Algumas doenças são monogênicas, ou seja, doenças cuja mutação causadora está presente em um único gene. São muito mais raras, e geralmente mais graves. Nestes casos, fatores ambientais têm pouca influência. Quando diagnosticada uma doença monogênica, isso não indica apenas uma propensão a desenvolver a doença, mas sim que o indivíduo, em geral, apresenta um risco alto de ser afetado ou gerar descendentes afetados.

Para casos assintomáticos, ou seja, doenças genéticas de manifestação tardia, o diagnóstico feito pela análise do DNA dá ao indivíduo portador de patologia com tal característica a oportunidade de decidir ter ou não filhos, evitando a transmissão do gene. A doença de Huntington é um exemplo onde os sintomas são manifestados, em geral, por volta da terceira ou quarta década de vida. O portador da doença de Huntington possui uma condição degenerativa que causa demência progressiva e morte. Pode ser autossômica dominante, levando a uma probabilidade de 50% de transmissão aos filhos do portador desta doença.

Teste de triagem populacional visam detectar os portadores de genes deletérios antes do aparecimento dos sintomas da doença ou do nascimento de uma criança afetada na família. Doenças como Tay Sachs e anemia falciforme, em judeus Aschkenazis e afrodescendentes respectivamente, são relativamente alta nestas populações (FARAH, 2007).

Uma família que possui um indivíduo afetado por uma doença genética deve receber informações referentes aos riscos recorrentes, prognósticos e possíveis tratamentos. Este conjunto de informações reportadas pelo geneticista clínico à família é chamada de aconselhamento genético.

2.2 Pipeline de Pré-processamento

No pré-processamento são realizados os processos de avaliação da qualidade, alinhamento e a anotação dos dados. Atualmente, já existem ferramentas que cumprem este papel, e estas foram utilizadas para isso.

O *pipeline* de pré-processamento inicia alinhando o fastq do paciente com um genoma referência. Uma ferramenta que pode ser usada para isso é a BWA. O resultado do mapeamento é um arquivo SAM/BAM que contém as informações de qualidade do alinhamento. Este arquivo é ordenado pelas coordenadas genômicas e então são marcadas as duplicatas de PCR. Uma ferramenta que pode ser usada para isto é a Picard. Em seguida, as variantes são encontradas segundo as recomendações de uma ferramenta como, por exemplo, a *Genome Analysis Tool Kit* (GATK), disponibilizando um arquivo no formato VCF, contendo a informação das mutações encontradas e da qualidade das variantes. Posteriormente, utilizando por exemplo, a ferramenta Snpeff é executada a anotação das variantes, ou seja, a identificação de qual o efeito dela na estrutura primária de uma proteína, por exemplo, identificando se ela é sinônimo ou não sinônimo. Outras informações são adicionadas, como a predição de alteração na estrutura tridimensional de uma proteína, conservação do aminoácido referência entre os organismos, e até mesmo a frequência de uma determinada mutação em população normal do projeto 1000 Genomas. Todo este conjunto de dados são salvos em um arquivo tabular, que posteriormente será usado como fonte para a análise. Neste projeto, este arquivo será a entrada para a ferramenta NGA (JOHNSON e colab., 2017).

2.3 Banco de dados biológicos

As sequências obtidas no teste genético não revelam nada sobre como essa informação genética controla o desenvolvimento de um organismo vivo. O mapeamento contra uma referência vai permitir identificar as mutações e o processo chamado de anotação do genoma é que tenta definir os genes (tanto os que codificam como os que não codificam proteínas) em um genoma e atribuir um papel para cada um (JOHNSON e colab., 2017).

Catalogar as mutações de pessoas em volta do mundo e descobrir as que aumentam o risco de várias doenças, tornaram possível determinar as sequências patogênicas. Estas informações em geral são organizadas em o que chamamos de banco de dados biológicos.

Banco de dados é uma coleção de dados de qualquer tamanho e complexidade, organizados e relacionados entre si usando um computador (ELMASRI, 2010). Por exemplo,

nomes e números de telefones de pessoas, compondo uma agenda. Um banco de dados computadorizado pode ser criado e mantido por programas construídos para esta tarefa ou por um sistema de software de uso geral que facilita o processo de criar e manter o banco de dados. Este sistema é chamado de Sistema Gerenciador de Banco de Dados (SGBD).

Os dados biológicos são altamente complexos quando comparados com a maioria dos outros domínios ou aplicações. Os esquemas nos bancos de dados biológicos mudam rapidamente. A formulação de consultas complexas é extremamente importante para os biólogos que irão utilizar estes dados, pois a maioria provavelmente não irá conhecer a estrutura do banco de dados (ELMASRI, 2010).

A seguir são descritos alguns bancos de dados biológicos.

2.3.1 Omim

Herança mendeliana do ser humano (*Mendelian Inheritance in Man* – OMIM) é uma base de conhecimento detalhado dos genes humanos e distúrbios genéticos. Consiste em resumos de texto completo de genes e fenótipos genéticos, particularmente desordens, e é útil para estudantes, pesquisadores e clínicos. Ele foi iniciado na década de 1960 como uma trilogia de catálogos de autossômica dominante, autossômica recessiva, e fenótipos ligados ao X (MCKUSICK, 2007).

Em 1987 foi publicado na internet, com o nome de OMIM. Desde dezembro de 1995, passou a ser distribuído pelo Centro Nacional de Informações sobre Biotecnologia (NCBI), que consiste em uma versão informatizada e disponível na web, atualizada diariamente, do livro *Mendelian Inheritance in Man*, editado pelo Dr. Victor MCKurisc, da Jonhns Hopkins University, sendo eles um catálogo de genes e doenças humanas herdadas de maneira mendeliana (KLUG e colab., 2010).

O OMIM é um banco de dados público, atualizado diariamente. Doenças genéticas causadas a partir de alterações cromossômicas maiores, não são incluídas neste banco (KLUG e colab., 2010).

O OMIM pode ser acessado nos sites <http://www.omim.org/> ou <http://www.ncbi.nlm.nih.gov/omim>.

2.3.2 Orphanet

O Orphanet foi criado na França pelo Instituto Nacional Francês para a Saúde e Investigação Médica em 1997. Esta iniciativa foi a semente para um esforço europeu que, a

partir de 2000, apoiado por subvenções da Comissão Europeia, tem vindo gradualmente a tornar-se um Consórcio de 40 países, na Europa e em todo o mundo (“Orphanet”, 1999).

O Orphanet reúne informações e promove o conhecimento sobre as doenças raras. O Orphanet mantém a nomenclatura de doenças raras e atribui número ORPHA, essencial para uma melhoria da visibilidade das doenças raras nos sistemas de informação em saúde e na investigação clínica.

O uso do número Orpha permite por exemplo, obter informações de uma determinada doença de forma simples. Por exemplo, sabendo que a doença Familiar colorectal cancer Type X possui o orpha número 440437 ao acessar a URL https://www.orpha.net/consor/cgi-bin/OC_Exp.php?Expert=440437 que contém o número Orpha no final, todos os dados do Orphanet sobre esta doença é exibido.

O Orphanet possui atualmente em sua base dados sobre 5856 doenças e 3573 genes cadastrados e pode ser acessado no site <https://www.orpha.net/>.

2.3.3 HPO

Human Phenotype Ontology (HPO) fornece recursos de bioinformática para a análise de doenças e fenótipos humanos. Foi inicialmente publicado em 2008 com o objetivo de permitir a integração de informações fenotípicas entre campos científicos e bancos de dados (KÖHLER e colab., 2017).

O HPO visa fornecer um vocabulário padronizado de anormalidades fenotípicas encontradas em doenças humanas. Cada termo no HPO descreve uma anormalidade fenotípica. Atualmente, a HPO contém aproximadamente 11.000 termos (ainda em crescimento) e mais de 115.000 anotações para doenças hereditárias. O HPO também fornece um grande conjunto de anotações de HPO para aproximadamente 4000 doenças comuns (KÖHLER e colab., 2017).

Todos os dados do HPO estão disponíveis para download no site <http://human-phenotype-ontology.github.io/downloads.html> ou podem ser acessados *on-line* pelo site <http://human-phenotype-ontology.github.io/tools.html>.

2.3.4 Hugo

O *Human Genome Organisation* (HUGO) é responsável pela aprovação de símbolos e nomes únicos, para genes de codificação de proteínas, genes de ncRNA e pseudogenes humanos, para permitir comunicação científica inequívoca (YATES e colab., 2017).

O HUGO possui atualmente um total de 41361 símbolos cadastrado e pode ser acessado através da URL <https://www.genenames.org/>.

O HUGO permite o download de seus dados no formato txt (separado por tab) e json.

2.3.5 1000 Genomas

O projeto 1000 Genomas (AUTON e colab., 2015) foi estabelecido para fornecer uma descrição detalhada da variação genética humana a partir do sequenciamento de genoma completo de um conjunto de diversificado de 2504 indivíduos de 26 populações.

Os dados do projeto estão disponíveis de forma pública na internet no site <http://www.1000genomes.org/data#download>. Comparando as variantes identificadas de indivíduos normais do Projeto 1000 Genomas com os dados do paciente é possível identificar e variantes que não são patogênicas, auxiliando na clínica média.

2.3.6 ICD

Uma das terminologias mais conhecidas para classificação de doenças e suas modificações clínicas é a *International Classification of Diseases* (ICD). Publicado pela primeira vez em 1893, foi revisado em intervalos aproximados de 10 anos, primeiro pelo Instituto Internacional de Estatística e, mais tarde, pela Organização Mundial da Saúde (OMS). A décima edição foi publicada em 1992 (SHORTLIFFE e CIMINO, 2015). Atualmente a ICD encontra-se na sua versão 11 (“WHO | ICD-11 Revision”, 2018).

2.4 Formatos de arquivos

Muitos dos bancos de dados biológicos populares começaram como bancos de dados de arquivos simples e, devido a este legado, muitos programas de bioinformática têm requisitos regidos no formato dos dados de entrada.

A seguir serão apresentados alguns formatos de arquivos normalmente utilizados na bioinformática.

2.4.1 VCF

O *Variant Call Format* (VCF) é um formato de arquivo de texto, utilizado para armazenar informações de variantes genéticas.

A especificação do VCF não é mais mantida pelo Projeto 1000 Genomes. O grupo que lidera a gestão e expansão do formato é a *Global Alliance for Genomics e Health Data Working group file format team* (SAMTOOLS, 2015).

Na versão 4.3 a codificação de caracteres dos arquivos VCF é UTF-8. Em suas primeiras linhas de dados, o VCF pode conter meta-informações, que são iniciadas com dois símbolos “#”; as meta-informações são opcionais, mas caso sejam informadas devem estar no formato chave=valor. Após as meta-informações, o VCF possui uma linha de cabeçalho que inicia com o símbolo “#”, esta linha irá descrever os campos das linhas de dados. Após o cabeçalho o VCF possui as linhas de dados, onde cada linha representa uma informação e seus campos são separados por *tabs*.

Os dados possuem oito campos obrigatórios: CHROM (cromossomo ao qual pertence a variante), POS (a posição da variante em relação ao cromossomo de referência), ID (identificador, caso possua mais de um, os identificadores devem ser separados por ponto e vírgula, caso esteja usando um identificador do dbSNP o prefixo “rs” deve ser usado antes do número, caso não possua um identificador disponível, o campo pode ficar com um *string* vazia ou usar o símbolo de ponto e vírgula) REF (a base ou o conjunto de bases de referência para a variante), ALT (a base ou o conjunto de bases alteradas na variante), QUAL (qualidade, um *score* de confiança da alteração da variante), FILTER (estado do processo de filtragem na montagem do arquivo VCF), INFO (informação adicional). Linhas em branco não são permitidas no VCF. Caso haja alguma linha em branco, um ponto (“.”) deve ser usado no lugar.

Para garantir a interoperabilidade entre plataformas, as implementações compatíveis com VCF devem suportar as convenções de novas linhas por LF (\n) e CR+LF (\r\n) (SAMTOOLS, 2015).

Na Figura 3 pode ser visto um exemplo de arquivo VCF.

Figura 3 - Exemplo de arquivo VCF

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0/0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

Fonte: (SAM e colab., 2018)

2.4.2 SAM / BAM

O arquivo stands for Sequence Alignment/Map format (SAM) é um arquivo de texto, com suas colunas separados por *tab*. O arquivo BAM é similar ao SAM, diferenciando por armazenar os dados em um formato binário que ocupam menos espaço em disco.

O SAM opcionalmente possui um cabeçalho que deve iniciar pelo símbolo “@”. Cada linha possui 11 colunas obrigatórias, colunas estas que possuem informações essenciais de alinhamento, estes campos podem ser vistos na Tabela 8.

Tabela 1 - Campos do arquivo SAM

Coluna	Campo	Descrição
01	QNAME	Nome da sequência
02	FLAG	Bits indicando informações sobre o alinhamento
03	RNAME	Nome da sequência na referência
04	POS	Posição mais à esquerda da sequência que se alinha na referência
05	MAPQ	Qualidade do Mapeamento
06	CIGAR	Informações sobre o alinhamento
07	RNEXT	Referência da próxima sequência no par/segmento
08	PNEXT	Posição da próxima sequência no par/segmento
09	TLEN	Tamanho do modelo observado
10	SEQ	Sequência de nucleotídeos (sequência)
11	QUAL	Qualidade da sequência codificada

É um formato de texto delimitado por TAB que consiste em uma seção de cabeçalho, que é opcional, e uma seção de alinhamento. Se presente, o cabeçalho deve ser anterior aos alinhamentos.

As linhas de cabeçalho começam com "@", enquanto as linhas de alinhamento não. Cada linha de alinhamento possui 11 campos obrigatórios para informações essenciais de alinhamento. Na Figura 4 pode ser visto um exemplo de arquivo SAM (SAM e colab., 2018).

Figura 4 - Exemplo de arquivo SAM

```
@HD VN:1.6 S0:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Fonte: (SAM e colab., 2018)

2.4.3 Fasta

O formato de arquivo Fasta é um formato baseado em texto, utilizado para armazenar sequências de nucleotídeos, onde a primeira linha inicia com o símbolo de maior ">" seguido por breve comentário informado a descrição da sequência. A segunda linha é onde fica a sequência de nucleotídeos propriamente dita.

Na Figura 5 pode ser visualizado um exemplo de um pequeno trecho de um arquivo Fasta. Neste exemplo está sendo apresentado apenas uma sequência, mas o arquivo Fasta pode conter diversas sequências.

Figura 5 - Arquivo Fasta

```

>J02459.1 Enterobacteria phage lambda, complete genome
GGGCGGCGACCTCGCGGGTTTTTCGCTATTTATGAAAATTTTCCGGTTTAAGGCGTTTCCGTTCTTCTTCG
TCATAACTTAATGTTTTTATTTAAAATACCCTCTGAAAAGAAAGGAAACGACAGGTGCTGAAAGCGAGGC
TTTTTGGCCTCTGTCGTTTCCTTTCTCTGTTTTGTCCGTGGAATGAACAATGGAAGTCAACAAAAGCA
GCTGGCTGACATTTTCGGTGCGAGTATCCGTACCATTGAGAACTGGCAGGAACAGGGAATGCCCGTTCTG
CGAGGCGGTGGCAAGGGTAATGAGGTGCTTTATGACTCTGCCGCCGTGTCATAAAATGGTATGCCGAAAGGG
ATGCTGAAATTGAGAACGAAAAGCTGCGCCGGGAGGTTGAAGAACTGCGGCAGGCCAGCGAGGCAGATCT
CCAGCCAGGAACATTGAGTACGAACGCCATCGACTTACGCGTGCGCAGGCCGACGCACAGGAAGTGAAG
AATGCCAGAGACTCCGCTGAAGTGGTGGAAACCGCATTCTGTACTTTCGTGCTGTGCGGGATCGCAGGTG

```

Fonte: (SALGADO e colab., 2016)

2.4.4 Fastq

O Fastq é um formato de arquivo para armazenar em conjunto os dados de sequenciamento e a qualidade associada a cada base. Tanto o nucleotídeo quanto a qualidade são representados por um carácter ASCII cada. O Fastq é uma extensão do formato de arquivo Fasta, se diferenciando por armazenar a um valor numérico que é a qualidade associada com cada nucleotídeo em uma sequência.

Cada sequência normalmente utiliza quatro linhas, onde a primeira inicia com o caractere arroba “@” seguido do identificador e opcionalmente uma descrição da sequência, a segunda linha é a sequência propriamente dita. A terceira linha contém o caractere símbolo da adição “+” onde opcionalmente, após o símbolo repete o identificador da sequência, a quarta e última linha contém a qualidade, referente a nucleotídeo de mesma posição da sequência.

Na Figura 6 pode ser visualizado um pedaço de um arquivo no formato Fastq.

Figura 6 - Exemplo de arquivo fastq

```
@ERR000916.1 IL21_771:6:1:884:502/1
GAAATATGGAATGCAAAGTATATGGCACAGCTGATT
+
????????????;???;????????;??>?:;9;==5
@ERR000916.2 IL21_771:6:1:887:543/1
GAAACTATATCTAAATAATTATGGTAAGTATGAGAA
+
????????????<????????????;<?<:??<=6:2:++
@ERR000916.3 IL21_771:6:1:878:213/1
GGAGTACTCACCACAGGACTGGGGTGGTGTATGGCCA
+
????????????????@??;?????????@?@?;=;=;;
@ERR000916.4 IL21_771:6:1:877:163/1
GAAACTCCTTGAGAACAGGGATTACTTCATTGGTCT
+
????????????????@?@?@;??;@>?@?==2;;;
```

Fonte: (PARK e colab., 2017)

2.5 Ferramentas para alinhamento e anotação

Após sequenciamento, é importante iniciar a análises acrescentando informações que visam melhorar o conhecimento biológico dos organismos. Uma abordagem para se obter estas informações é mapear as sequências contra uma referência, selecionar apenas as que possuem alguma mutação e anotar estas mutações.

A seguir será apresentado algumas ferramentas para essas atividades, incluído também ferramenta para a conversão no formato de arquivos.

2.5.1 BWA

O Burrows-Wheeler Aligner (BWA) é um pacote de software para o mapeamento de sequências contra um genoma referência. É composto de três algoritmos BWA-backtrack, BWA-SW e BWA-MEM.

O BWA-backtrack é projetado para leitura de sequência Illumine até 100bp, enquanto os dois restantes variaram de 70pb a 1Mpp. O BWA-MEM e o mais recente e geralmente é o recomendado para consultas de alta qualidade, uma vez que ele é o mais rápido e preciso.

A entrada é um arquivo Fastq e saída é um arquivo SAM.

2.5.2 Picard

O Picard é um programa de linha de comando que permite a manipulação de arquivos SAM/BAM CRAM e VCF. O Picard permite indexar arquivos SAM/BAM e remover duplicatas de PCR, ou seja, remover as leituras duplicadas.

Por ser desenvolvido em Java o Picard é multiplataforma e está disponível para download no site <http://broadinstitute.github.io/picard/>.

Uma vez que a saída do BWA é um arquivo SAM e o GATK necessita de um arquivo BAM como entrada, o Picard é utilizado para fazer esta conversão de formato.

2.5.3 GATK

Genome Analysis Toolkit (GATK) é um kit de ferramentas para análise do genoma, com foco na análise de variantes, genotipagem e garantia da qualidade dos dados. O GATK não prove uma interface gráfica, sua utilização se dá através de linha de comando. Executando o programa GenomeAnalysisTk.jar seguido dos parâmetros desejado o usuário invoca o GATK e pelos parâmetros informa qual função do mesmo deseja utilizar (SCHMIDT, 2009).

A entrada do GATK é um arquivo BAM e a saída é um arquivo VCF.

2.5.4 snpEFF

O SnpEff permite a predição do efeito causado pelas variantes com base em sua localização genômica (CINGOLANI et al., 2012).

A entrada do SnpEFF são as variantes as quais se deseja prever seu efeito, normalmente em um arquivo no formato VCF. O SnpEff anota as variantes e seu efeito que pode ser: high, moderate, low ou modifier (CINGOLANI et al., 2015). A saída do snpEFF é um arquivo VCF, que irá conter as mutações anotadas.

2.5.5 VEP

O Variant Effect Predictor (VEP) determina o efeito de em genes, transcritos e sequência de proteínas, bem como regiões reguladoras (MCLAREN e colab., 2016).

Com base na coordenada da variante e as alterações do nucleotídeo, o VEP informa: gene e transcrito afetados pelas variantes; Consequência das variantes na sequência de proteínas (por exemplo, *stop gained*, *missense*, *stop lost*, *frameshift*); *Score* para o SIFT e PolyPhen para alterações na sequência proteica etc.

O VEP pode ser acessado na web através da url <https://www.ensembl.org/Tools/VEP>, ou por meio de API REST <https://rest.ensembl.org/#VEP> ou mesmo por scripts Perl disponíveis para download em <https://github.com/Ensembl/ensembl-vep> que podem ser instalados local.

O VEP realiza a mesma função do snpEFF, possui como entrada e saída, também um arquivo VCF.

2.5.6 UMD Predictor

O UMD-Predictor está disponível para download no site <http://umd-predictor.eu>, ele é utilizado para anotar substituições de *complementary* DNA (cDNA) de todos os transcritos humanos quanto à sua potencial patogenicidade. Ele combina propriedades bioquímicas, impacto em sinais de *splicing*, localização em domínios de proteínas, frequência de variação na população global e conservação através da matriz de substituição global BLOSUM62 e uma conservação específica de proteína entre 100 espécies (SALGADO e colab., 2016).

3 TRABALHOS CORRELATOS

Existem alguns trabalhos que propõem modelos para análise de exoma humano. A maioria das ferramentas disponíveis atualmente foram desenvolvidos para fins de pesquisa exploratória e muitas vezes não fornecer uma solução completa e eficiente na pesquisa e atendimento clínico (PANDEY et al., 2016). Muitos laboratórios ainda utilizam planilha eletrônica para gerenciar conjuntos de dados de variantes, a despeito de este tipo de ferramenta não ser escalável e sujeita a erros (WANG et al., 2015).

Esta seção apresenta três trabalhos que propõem auxiliar a análise de exoma humano destacando suas funcionalidades e limitações.

3.1 ClinLabGeneticist

A ferramenta ClinLabGeneticist implementa um fluxo de trabalho para a gestão de variantes e diagnósticos de doenças. Nesta ferramenta, todo o fluxo de trabalho é integrado em uma plataforma única de gerenciamento de dados. Além de ter uma interface gráfica, existem *hiperlinks* para bases de dados externos, o que fornece uma conveniência para os utilizadores (WANG e colab., 2015).

A ClinLabGeneticist possui painéis de controle, um para os revisores e outro para os administradores, e suas funcionalidades pode ser vista na *Figura 7*.

Várias plataformas de sequenciamento incluindo Ion Torrent, Agilent, NimbleGen são suportadas. O arquivo de entrada deve possuir as seguintes colunas: número de cromossomos, coordenar cromossomo, alelo referência, alelo amostra, símbolo gene, transcrito ID, alteração de nucleotídeos, alteração de aminoácidos, SIFT por predição funcional, PolyPhen-2 predição funcional, phyloP p-value, dbSNP ID e 1000 frequência do alelo genoma.

O sistema é projetado para salvar automaticamente as anotações a cada 30 segundos e conta com *hiperlinks* para auxiliar a anotações das variantes, ligando às suas bases de dados externas correspondentes (por exemplo dbSNP, OMIM, ClinVar etc.) sobre a qual a anotação se baseia. Além disso, a localização do cromossomo possui *link* para a página UCSC, o símbolo do gene possui *link* para o GeneCards para uma descrição mais detalhada do gene, e cada gene é ligado a NCBI PubMed para literatura relevante. Embora o ClinLabGeneticist utilize informações de diversos bancos de dados, ela não permite a incorporação de novos bancos de dados por parte do usuário, o que impede que um usuário que queira utilizar um banco de dados

específico que não esta lista utilize este banco, um recurso que é especialmente importante para pesquisa que pode ter a necessidade de uso de base de conhecimentos diferente das bases utilizadas rotineiramente.

Figura 7 - A arquitetura e as funcionalidades do ClinLabGeneticist

Fonte: (WANG e colab., 2015)

Cada variante pode ser atribuída a pelo menos dois revisores independentes, onde caso haja interpretações discordantes sobre a mesma variante, definem-se reuniões de grupo com os revisores para resolver as discrepâncias. A resolução de interpretações discordantes ser feita por meio de reuniões fora do sistema, pode gerar um atraso na análise, além disso o fato de estas avaliações serem feitas fora do sistema, em um futuro pode se perder os motivos que levaram a uma definição ou outra.

O sistema é gerenciado por um administrador de sistema cujas responsabilidades incluem a concessão de privilégios, adicionando ou removendo colaboradores e gestão de arquivos variantes.

As anotações funcionais das variantes, tais como dbNSFP e resultados pré-processamento de variantes genéticas conhecidas geradas pelas ferramentas como SIFT, PolyPhen, ANNOVAR, SnpEff e MutationAssessor são exibidas.

ClinLabGeneticist é construído sobre a plataforma Windows (Windows 7 e 8). Arquitetura convencional de cliente/servidor é utilizada para apoiar concorrência e multiusuários. O sistema operacional Windows é o utilizado na cliente, o servidor que executa

o SGBD MySQL está implantado em Linux. A necessidade da implantação em plataforma Windows, dificulta a utilização de possíveis usuário, que utilizam outras plataformas como Linux e Mac ou mesmo usuários que utilizam a plataforma Windows, mas com usuários sem privilégios para instalação.

3.2 VariantStudio

O software VariantStudio possui recursos para anotar, filtrar, exportar e interpretar dados, permitindo que os pesquisadores explorem, identifiquem e classifiquem as variantes relevantes para a doença e relatem resultados biologicamente significativos. Na *Figura 8* é exibido o fluxo de trabalho do VariantStudio (ILLUMINA, 2016).

Figura 8 - fluxo de trabalho do VariantStudio

Fonte: (ILLUMINA, 2016)

O software VEP (descrito em 2.5.5) é o recurso utilizado para o processo de anotação, onde o VarianteStudio consolida informações de várias origens em um único banco de dados simplificando o processo de anotações nos níveis de variantes, genes e transcrição. Na *Figura 9* é possível visualizar a tela principal do VariantStudio.

Figura 9 - Illumina VariantStudio

Gene	Variant	Chr	Coordinate	Classification	Type	Genotype	Exonic	Filters	Quality	GQX	Inherited From	Alt Variant Freq	Read Depth	Alt Read Depth	Allele Depths
MSH6	AC>AC/A	2	48030639	...	deletion	het	yes	PASS	100	100		3.01	27546	828	26718...
MSH6	A>A/JAC	2	48030639	...	insertion	het	yes	PASS	100	100		46.7	27546	12863	14683...
MSH6	A>A/ACC	2	48030639	...	insertion	het	yes	PASS	100	100		4.52	27546	1246	26300...
CTNBN1	C>C/A	3	41266101	...	snv	het	yes	PASS	100	100		35.81	21474	7690	13757...
PDGFRA	A>G/G	4	55141055	...	snv	hom	yes	PASS	100	100		99.62	17756	17688	58,17...
FBXW7	TC>TC/T	4	153244155	...	deletion	het	yes	PASS	100	100		56.03	10342	5795	4547...
EGFR	G>G/A	7	55241707	...	snv	het	yes	PASS	100	100		32.31	18483	5971	12499...
GNAQ	GAAA>...	9	80343587	...	deletion	het	yes	R8	100	100		15.65	35325	5530	29795...
GNAQ	GAA>G...	9	80343587	...	deletion	het	yes	R8	100	100		41.31	35325	14593	20732...
GNAQ	GAA>GAA/G	9	80343587	...	deletion	het	yes	R8	100	100		28.57	35325	10094	25231...
GNAQ	GA>GA/G	9	80343587	...	deletion	het	yes	R8	100	100		6.83	35325	2413	32912...
KRAS	A>A/G	12	25362777	...	snv	het	yes	PASS	100	100		47.66	20135	9597	10517...
KRAS	C>C/G	12	25398285	...	snv	het	yes	PASS	100	100		5.23	17492	914	16540...
MAP2K1	A>A/C	15	66727451	...	snv	het	yes	PASS	100	100		48.57	12141	5897	6190...
TP53	G>C/C	17	7579472	...	snv	hom	yes	PASS	100	100		99.5	22445	22332	47,22...

Fonte: (ILLUMINA, 2016)

Informações sobre a associação de doenças conhecidas podem ser obtidas no Catálogo de Mutações Somáticas no Câncer (COSMIC), ClinVar, e OMIM, algoritmos como o Polimorfismo Phenotyping (PolyPhen) e SIFT suportam os processos de anotação que permite que os pesquisadores identifiquem a importância biológica em dados variantes, porém não possui *links* para as páginas destes banco, onde o usuário poderia consultar informações adicionais.

O sequenciamento do genoma total e do exoma detecta muitas vezes muitas variantes por amostra, exigindo uma análise mais aprofundada para identificar as poucas variantes associadas a um determinado fenótipo.

Os pesquisadores podem isolar as principais variantes que são consequentes ao fenótipo em estudo utilizando opções de filtragem. O VariantStudio oferece filtros comumente aplicados, incluindo qualidade variante, frequência, impacto funcional e associação de doenças conhecidas, que são facilmente acessíveis através da interface do usuário, mas não possui um filtro pré-definido, desta forma a utilização do VariantStudio tem uma dependência muito grande do conhecimento do usuário, não apoiando aqueles que não tem muita experiência neste tipo de análise.

Filtros com base em qualquer informação disponível, incluindo anotações

personalizadas importadas pelo usuário é suportada.

É oferecido a opção de construir fluxos de trabalho salvando os filtros aplicados, permitindo que os pesquisadores padronizem fluxos de trabalho e simplifiquem a análise.

Além da filtragem de amostra única, o software VariantStudio permite comparações de vários exemplos o auxilia na identificação de variantes causadoras. Mutações somáticas podem ser identificadas em pares tumor-normais por filtração de variantes que estão presentes em uma amostra de tumor, mas ausentes em sua contraparte normal.

O VariantStudio é um software *standalone*, que executa na plataforma Windows e é comercializado pela Illumina.

3.3 Mendel,MD

O software Mendel,MD foi desenvolvido para investigação dos casos clínicos recebidos pelo laboratório de genômica clínica da Faculdade de Medicina da Universidade Federal de Minas Gerais, Brasil. O Mendel,MD permite o armazenamento, anotação e seleção de variantes dos pacientes (GUIMARÃES e colab., 2015).

A ferramenta é acessada via navegador Web e está disponível no site <http://mendel.medicina.ufmg.br/> através do usuário gregor e senha mendel,md. A entrada dos dados se faz através do upload de um arquivo VCF, que é anotado e posteriormente persistido no banco de dados.

A interface permite visualizar diversas informações sobre cada indivíduo como o nome do arquivo, o número de variantes, a data de envio, o estado atual do arquivo no sistema (*Figura 10*).

A anotação das variantes é realizada integrando diversas ferramentas e fontes de informações, sendo a SnpEff uma das mais importantes que classifica a variante de acordo com seu impacto.

Os dados de doenças foram obtidos do OMIM, o que permitiu a busca de genes associados a uma determinada doença, sendo este um passo importante na seleção das variantes. Embora o OMIM seja útil para diversas análises o uso exclusivo desta base de dados pode ser um limitador na utilização desta ferramenta para diversas outras análises.

A seleção de variantes ocorre em três etapas através de um formulário web que permite a eliminação de variantes utilizando diferentes critérios de filtragem para tentar identificar a variante que possa ser responsável por causa de doença no paciente.

Figura 10 - Visualização do pacientes no Mendel,MD

257	mm150442ys	Edit Browse Delete	sdpena	38618	July 1, 2015, 1:57 p.m.	July 2, 2015, 7:23 a.m.	0:19:47.130665	2:04:44.160600	None	populated	Reannotate Individual Repopulate Individual Populate to MongoDB
256	mm150524ao	Edit Browse Delete	sdpena	101054	July 1, 2015, 1:49 p.m.	July 2, 2015, 9:29 a.m.	0:35:06.568579	4:10:29.182620	None	populated	Reannotate Individual Repopulate Individual Populate to MongoDB
255	mm150527wf	Edit Browse Delete	sdpena	101754	July 1, 2015, 1:42 p.m.	July 2, 2015, 9:26 a.m.	0:32:11.318878	4:07:11.888756	None	populated	Reannotate Individual Repopulate Individual Populate to MongoDB
254	mm150531am	Edit Browse Delete	sdpena	98474	July 1, 2015, 1:04 p.m.	July 2, 2015, 5:18 a.m.	0:21:28.254705	5:01:34.938282	None	populated	Reannotate Individual Repopulate Individual Populate to MongoDB
253	mm150539ps	Edit Browse Delete	sdpena	101619	July 1, 2015, 12:54 p.m.	July 2, 2015, 5:18 a.m.	0:23:35.855933	8:05:02.637177	None	populated	Reannotate Individual Repopulate Individual Populate to MongoDB
252	mm150544si	Edit Browse Delete	sdpena	97366	July 1, 2015, 12:43 p.m.	July 2, 2015, 5:18 a.m.	0:23:01.533086	8:05:03.016698	None	populated	Reannotate Individual Repopulate Individual Populate to MongoDB

Fonte: (GUIMARÃES e colab., 2015)

Na primeira etapa é selecionado os pacientes da análise e os pacientes que serão utilizados no controle, opcionalmente uma lista de genes e SNPs que podem ser para inclusão ou exclusão também podem ser informadas nesta etapa (*Figura 11*).

Figura 11 - Primeira etapa na seleção de variantes do Mendel,MD

+ Filter Options

MAIN
VARIANTS
DATABASES

SELECT VARIANTS FROM

EXCLUDE VARIANTS FROM

INDIVIDUALS:
SELECT YOUR CASES

GROUPS:
SELECT YOUR GROUPS

EXCLUDE INDIVIDUALS:
SELECT YOUR CONTRO

EXCLUDE GROUPS:
SELECT YOUR GROUPS

SNP LIST:

SAVED GENE LIST:
SELECT YOUR GENE LI:

EXCLUDE SNP LIST:

EXCLUDE SAVED GENE LIST:
SELECT YOUR GENE LI:

GENE LIST:

EXCLUDE GENE LIST:

SELECT INHERITANCE:

RECESSIVE HOMOZYGOUS
RECESSIVE COMPOUND HETEROZYGOUS
DOMINANT HETEROZYGOUS
X-LINKED RECESSIVE HEMIZYGOUS
X-LINKED DOMINANT HETEROZYGOUS

SELECT YOUR DISEASES:

OMIM:

CLINICAL GENOMICS DATABASE:

HGMD:

☐ OPEN RESULT IN A NEW WINDOW

SUBMIT

RESET FILTER

Fonte: (GUIMARÃES e colab., 2015)

A segunda etapa o usuário possui diversas opções para selecionar as variantes, como por exemplo, se a variante é homozigótica ou heterozigótica, por cromossomo, pelos dados da anotação (efeito, impacto, qualidade etc.).

Na terceira etapa são informados os valores mínimos e máximos de frequência, outros dados da anotação como escores SIFT e Polyphen-2. Além disso é possível selecionar genes relacionados a doenças específicas do OMIM.

A ferramenta possui também uma opção de filtro padrão, a configuração deste filtro aplica as seguintes restrições: SnpEff Impact: *high* ou *moderate*; Profundidade de Leitura maior ou igual a 10; mostrar apenas variantes em genes comuns aos indivíduos selecionados; Excluir Variantes presentes no banco VariSNP; Frequência da variante no 1000Genomes menor que 0.005; Frequência da variante no dbSNP137 menor que 0.005; Frequência da variante no Exome Variant Server menor que 0.005.

As variantes de interesse podem ser visualizadas e/ou exportadas para um arquivo cvs para serem investigadas manualmente por um clínico utilizando um *software* de planilha eletrônica. Por entender que o usuário pode e deve realizar toda a seleção e investigação na ferramenta.

Comparação de exomas de diferentes indivíduos e até mesmo do mesmo indivíduo gerados a partir de tecnologias diferentes também é suportado. Esta análise é realizada em duas etapas, onde na primeira é selecionado apenas as posições comuns aos dois e em seguida é verificado se o genótipo nessas duas posições é igual ou diferente.

3.4 Comparativo

Embora as ferramentas apresentadas tenham grande valia, sua utilização apresenta limitações no uso de grupos que trabalhe tanto com pesquisa como atendimento clínico e necessita de facilidade na comunicação entre diversos membros do grupo.

O NGA implementa uma fluxo de trabalho que permite a integração entre membros da equipe, para apoiar a análise, este fluxo de trabalho presente no NGA não está implementado no VariantStudio e Mendel,MD. Embora o ClinLabGeneticist implemente o fluxo de trabalho, em caso de discordância na análise a ferramenta não oferece uma forma simples de solucionar este tipo de impasse. A NGA permite que ao classificar uma variante o revisor indique seus motivos incluindo referências a artigos do motivo que o levou a tal classificado, desta forma caso haja discordância o gestor consegue consultar estas informações e classificar a mesma.

Realizar o acesso ao sistema por um navegador web, possibilita acesso ao sistema

intendente da arquitetura que o cliente esteja utilizando, necessitando apenas de um navegador web. Esta possibilidade de acesso por um navegador web está presente apenas no NGA e Mendel,MD. Para o ClinLabGeneticist e VariantStudio é necessário instalar um aplicativo na máquina do utilizador que precisar atender aos requisitos de hardware e software.

O NGA assim como o Mendel,MD permite incorporar novas informações de banco de dados biológico, garantido que o usuário consiga manter o sistema atualizado e útil. O NGA permite também a incorporação de novos bancos e suas respectivas informações.

Ao classificar uma variante, o sistema está gerando uma nova informação que pode ser muito útil em novas análise, o NGA permite que estas novas informações geradas na própria ferramenta sejam utilizadas em novas análises.

A exportação dos dados presentes nas ferramentas VariantStudio e Mendel,MD não está presente no NGA, por entendermos que a ferramenta deve suprir as necessidades na análise, desta forma não há a necessidade de exportar os dados para utilização em outras ferramentas.

No *Quadro 1* é possível visualizar as funcionalidades presentes em cada uma das ferramentas apresentada, juntamente com a NGA.

Quadro 1 - Comparativo entre os trabalhos correlatos e a NGA

(continua)

	ClinLabGeneticist	VariantStudio	Mendel,MD	NGA
Workflow	X			X
Interpretações discordantes são tratadas no sistema				X
Não necessita instalação nos clientes			X	X
Multiplataforma			X	X
O software inclui uma base de conhecimento que auxilia o usuário na seleção das variantes			X	X
Permite incluir a base de conhecimento de outros bancos de dados				X

Quadro 1 - Comparativo entre os trabalhos correlatos e a NGA

(conclusão)

	ClinLabGeneticist	VariantStudio	Mendel,MD	NGA
Permite a exploração das variantes na própria ferramenta	X	X		X
Permite a exportação dos dados		X	X	
Permite que o conhecimento gerado na ferramenta seja utilizado em novas análises				X
Auxilia a comunicação entre usuários envolvidos na análise	X			X
Existem hiperlinks para bases de dados externos, o que fornece uma conveniência para os utilizadores	X			X

4 PROPOSTA DO TRABALHO

O teste genético é realizado por meio da sequência de alguns procedimentos, como foi apresentado na *Figura 1*. Este capítulo apresenta a modelagem proposta para o processo de análise do sequenciamento de exoma humano. O modelo foi construído com base nos pré-requisitos estabelecidos junto aos membros do Centro de Medicina Genômica (CMG) do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP), com o objetivo de produzir um sistema computacional para seleção e classificação de variantes.

O procedimento de levantamento de requisitos permitiu que se chegasse a um modelo para o processo de análise, dividido em quatro fases, como mostrado na *Figura 12*: anotação, seleção das variantes, revisão das variantes e geração do laudo.

Figura 12 – Fases do modelo de análise de NGS

4.1 Anotação

O processo de anotação inicia-se no *pipeline* de pré-processamento; porém algumas informações que facilitam a seleção de variantes não são incluídas na fase de anotação do *pipeline*. Por este motivo, nesta proposta a fase de análise inclui outras anotações para facilitar a seleção das variantes.

Neste processo serão incluídas à variante, as doenças associadas à mesma. Estas informações são obtidas dos bancos de dados utilizados na análise que podem ser por exemplo: HPO, OMIM e Orphanet.

4.2 Seleção das variantes

A seleção das variantes representa o grande desafio da análise. Para tal as variantes anotadas são selecionadas por meio da aplicação de filtros.

Um método padrão de seleção – chamado aqui de **filtro padrão** - foi criado com base no conhecimento dos especialistas do CMG-HCFMRP, que participaram voluntariamente deste projeto. A ideia é que este deve ser o método ideal para análise de atendimento clínico, assim como para diversas análises de pesquisa. Em casos onde o filtro padrão não contemplar o objetivo do usuário, este tem a liberdade de aplicar modificações ao método de acordo com a sua necessidade, ou mesmo selecionar um filtro personalizado criado por ele ou por outro usuário.

Quando mais de um filtro é aplicado, as variantes selecionadas serão o resultado da intersecção dos filtros aplicados.

A seguir é descrito o filtro padrão que consiste na combinação de um conjunto de filtros.

4.2.1 Filtro padrão

O filtro padrão aplica cinco filtros: suspeita de diagnostico, efeito, recessivo, UMD e frequência.

O filtro por suspeita de diagnostico irá selecionar os genes relacionados a doença informada na análise.

No filtro por efeito, será selecionado as variantes com efeito *missence*, *splice* e todos os *inframes*.

Caso a doença seja recessiva, será aplicado o filtro de recessivo.

No filtro de umd será selecionado apenas as variantes patogênico ou provavelmente

patogênico.

O filtro pela frequência será aplicado, da seguinte forma: caso a doença seja rara, selecionam-se apenas as variantes presentes em menos de 10% da população, caso a doença não seja uma doença rara, serão selecionadas apenas as variantes presentes em mais de 10% da população.

A Figura 13 apresenta o fluxograma de filtros e decisões que compõem o algoritmo de aplicação do filtro padrão.

Figura 13 - Filtro padrão

4.2.2 Filtros

A seguir são descritos os filtros disponíveis na NGA.

Filtro por suspeita de diagnóstico

O filtro por suspeita de diagnóstico, utiliza as informações provenientes de bancos de dados biológicos públicos, previamente importado no sistema, que podem ser por exemplo o *Online Mendelian Inheritance in Man* (OMIM) e o COSMIC, que permite identificar quais genes estão associadas à patologia.

Ao aplicar este filtro é selecionado apenas os genes relacionados à patologia.

Filtro por gene

O filtro por gene permite que um ou mais símbolos de genes sejam informados, o conjunto resultado terá apenas variantes destes genes.

O símbolo é o nome curto dado ao gene para identificar ele. Uma dificuldade encontrada com na implementação do filtro por gene, foi o fato de um mesmo gene ter mais de um símbolo, exemplo o gene RPS7 possui os sinônimos ES7 e S7 (ANNILO e colab., 1995). Iniciativas como o HUGO realizam a aprovação de símbolos e nomes únicos, para genes de codificação de proteínas, genes de ncRNA e pseudogenes humanos, porém pode ocorrer de algum banco de dados biológicos ou mesmo que a anotação realizada no teste genético na fase de anotação não referenciam os genes pelos símbolos aprovados no HUGO, por este motivo foi necessário implementar um dicionário de sinônimo, que permite não duplicar informação para casos por exemplo de informar um novo símbolo para um gene e caso este gene seja sinônimo, garantindo com isso que consultas pelos símbolo aprovado no HUGO ou por um sinônimo retorne as informações corretas.

Filtro por efeito

O campo *effect* armazena o efeito que a mutação produz nos genes conhecidos, este efeito pode ser predito por exemplo pela ferramenta SnpEff. O filtro por efeito permite que o usuário informe os efeitos que deseja selecionar onde apenas as variantes que possuem os efeitos selecionados serão selecionadas.

Filtro por herança

Algumas doenças são recessiva e para estas doenças é conhecido que para a mutação causar algum impacto no paciente, a mutação precisa ser homozigoto, para atender este tipo de busca, o filtro por herança permite que o usuário informe o tipo de herança que deseja selecionar heterozigoto e/ou homozigoto e neste caso, é selecionado apenas as variantes com o padrão de herança informado.

Filtro por UMD predictor

A ferramenta UMD predictor executada no *pipeline* de pré-análise fornece uma abordagem combinatória, para identificar potenciais variações patogênicas, que associa os seguintes dados: localização dentro da proteína, conservação, propriedades bioquímicas dos resíduos mutantes e do tipo selvagem, e o impacto potencial da variação no mRNA. As

variantes podem ser classificadas em quatro classes: *Probable Polymorphism*, *Polymorphism*, *Probably Pathogenic* ou *Pathogenic*. Este filtro permite que um ou mais impactos classificados pelo UMD Predictor seja informado e apenas as variantes com as classificações informadas são selecionadas (SALGADO e colab., 2016).

Filtro por frequência

Patologias podem ter ou não uma frequência alta, sendo esperado que mutações frequentemente encontradas na população não sejam as mutações relacionadas a doenças raras e vice-versa. Para permitir a busca por variantes frequentes ou raras foi implementado o filtro pela frequência da variante que permite ao usuário informar dois valores, mínimo e máximo e desta forma selecionando apenas as variantes no intervalo informado.

Outros filtros

Além dos filtros descritos é possível realizar o filtro por qualquer informação da variante. Para os valores textuais, o usuário pode informar por quais valores deseja selecionar, para os valores numéricos o usuário pode informar dois valores, mínimo e máximo e será selecionado os valores neste intervalo.

Na Quadro 2 é possível consultar os tipos de filtros que podem ser aplicados aos campos.

Quadro 2 - Filtros disponíveis

(continua)

Filtro	Tipo	Descrição
Cromossomo	Catégorico	Seleciona os cromossomos informados
Posição	Intervalo	Seleciona as variantes no intervalo informado
Alelo da Referência	Catégorico	Seleciona os alelos informados da referência
Alelo da Variante	Catégorico	Seleciona os alelos informados da variante
Gene	Catégorico	Seleciona os genes informados
Umd Predictor	Catégorico	Seleciona as categorias do UMD predictor informado
Zygosity	Catégorico	Seleciona a variante por sua similaridade ou dissimilaridade do DNA entre cromossomos
Allelic depth	Intervalo	Seleciona o intervalo de alelos relatado por variantes para o depth

Quadro 2 - Filtros disponíveis

(continuação)

Filtro	Tipo	Descrição
Allelic deph2	Intervalo	Seleciona o intervalo de alelos relatado por variantes para o deph2
Filter	Catégorico	Seleciona variantes que passou pelos controles de qualidade
Hgvs C	Catégorico	Seleciona variantes pelas sequências codificadora do HGVS
Hgvs P	Catégorico	Seleciona as variantes pela sequência proteica do HGVS
DbSNP	Catégorico	Seleciona as variantes pelo identificador do DbSNP
Exon/Intro	Catégorico	Seleciona as variantes informando o número de exon/intron
Typef	Catégorico	Seleciona por genes e transcrição
Effect	Catégorico	Seleciona pelo termo de ontologia da variante
Impact	Catégorico	Seleciona pela estimativa do impacto da variante
Clinvar significance	Catégorico	Seleciona pelo significado clínico da variante
Clinvar disease	Catégorico	Seleciona pelo nome da doença associada a variante
Clinvar accession	Catégorico	Seleciona pelo identificador de acesso e versão atribuída do Clinvar à relação genótipo/fenótipo
Clinvar allele type	Catégorico	Seleciona pelo alelo referência ou alterado.
Clinvar allele origin	Catégorico	Seleciona pela origem do alelo da variante.
Sift	Catégorico	Seleciona pela classificação do Sift
Polyphen hdiv	Catégorico	Seleciona pela classificação do Polyphen2 baseado no hdiv
Polyphen hvar	Catégorico	Seleciona pela classificação do Polyphen2 baseado no hvar
Mutation taster	Catégorico	Seleciona pela predição do mutationTaster
LRT	Catégorico	Seleciona pela predição do annovar
Gerp rs score	Intervalo	Seleciona pelo intervalo de conservação de cada nucleotídeo no alinhamento de múltiplas espécies

Quadro 2 - Filtros disponíveis

(continuação)

Filtro	Tipo	Descrição
Gerp neutral rate	Intervalo	Seleciona pelo intervalo de posições de restrição evolucionária para a estimativa de taxa evolutiva de máxima verossimilhança
Feature	Categórico	Seleciona pela característica da variante
Ensembl	Categórico	Seleciona pelo identificador da variante no Ensembl
Vertebrate genomes conservation score	Intervalo	Seleciona pelo grau de conservação do sítio da mutação
Interpro domain	Categórico	Seleciona pela classificação no InterPro
Status	Categórico	Seleciona pelo status da variante
Genotype	Categórico	Seleciona pela confiança que a atribuição do genótipo está correta
Read depth	Intervalo	Seleciona pela profundidade de leitura aproximada
Allele mut fraction	Intervalo	Seleciona o intervalo de alelo na amostra que são mutações
Mean base quality	Intervalo	Seleciona pela pontuação do teste Z de classificação de Wilcoxon da base alterada vs qualidade da base de referência.
Variant type	Categórico	Seleciona pelo tipo da variante utilizado no dbSNP
Validate	Booleano	Seleciona por validação em outro experimento
Donor splice site	Booleano	Seleciona se a variante está dentro da região doadora de splicing do íntron
Acceptor splice site	Booleano	Seleciona se a variante está localizada em um splice site
Is mutation	Booleano	Seleciona se a mutação tem baixa frequência de citação em periódicos e em outras fontes confiáveis
European variant freq	Intervalo	Seleciona pela frequência da variante na população europeia
African variant freq	Intervalo	Seleciona pela frequência da variante na população africana

Quadro 2 - Filtros disponíveis

(conclusão)

Filtro	Tipo	Descrição
Assian variant freq	Intervalo	Seleciona pela frequência da variante na população asiática
American variant freq	Intervalo	Seleciona pela frequência da variante na população americana
Whole variant freq.	Intervalo	Seleciona pela frequência da variante em todas as populações

4.4 Revisão das variantes

O subconjunto de variantes selecionadas pelo gestor deve ser enviado a especialistas da doença, que deverão realizar consultas e continuar a análise, retornando à informação se estas variantes são ou não relacionadas à doença em estudo. Na ferramenta NGA, o papel destes especialistas é de revisor.

Para isso além de seu conhecimento o revisor deve consultar os bancos de dados biológicos e classificar cada variante em uma das seguintes opções: *benign*, *likely benign*, *uncertain significance*, *likely pathogenic*, *pathogenic*.

O revisor pode também incluir textos descritivos para cada variante, para justificar sua escolha.

4.5 Laudo

A partir da análise realizada pelos revisores, o gestor deverá informar uma das seguintes opções: *benign*, *likely benign*, *uncertain significance*, *likely pathogenic* ou *pathogenic*, para cada uma das variantes enviadas aos revisores. Para realizar esta classificação o gestor além de todas as ferramentas oferecidas ao revisor também terá acesso às classificações geradas pelos revisores e a todos as informações incluídas para cada variantes.

Além da classificação o gestor irá gerar também o laudo do teste genético.

5 NEXT-GEN ANALYSER

A NGA é a ferramenta construída que implementa o modelo proposto, sendo uma ferramenta que apoia no teste genético para dar suporte à análise, realizando parte do processo de anotação e os processos de seleção de variantes realizados pelo gestor, validação pelos revisores e laudo final do gestor, implementando tudo em um fluxo de trabalho e em ambientes específicos para cada ator.

O fluxo de trabalho pré-definido da NGA facilita a comunicação entre o gestor que é quem conduz a análise, e os revisores que são os especialistas que apoiam o gestor na *análise* (Figura 14).

Figura 14 - Fluxo de trabalho do NGA

Além do gestor e do revisor um terceiro perfil de usuário foi criado, o administrador.

Após a instalação do sistema e/ou sempre que for necessário o administrador deve gerenciar os usuários do sistema atribuindo permissões de administrador, gestor e/ou revisor além de também gerenciar os dados utilizados no NGA, como por exemplo as populações, genes, patologia etc.

O fluxo de uma análise no NGA inicia com o cadastro de um paciente e importação do arquivo de variantes resultante do *pipeline* de pré-processamento pelo gestor. Com as variantes importadas, o gestor pode então criar uma análise informando o paciente a patologia suspeita e outras informações.

Esta análise criada pode ou não utilizar o filtro padrão ou qualquer outro filtro existente,

onde, independentemente do filtro, o gestor tem a opção de personalizar a seleção de variantes, restringindo ou ampliando o escopo dos filtros.

Após a seleção das variantes, o gestor indica os revisores que irão apoiá-lo nesta análise.

O revisor recebe a lista de variantes selecionadas pelo gestor e classifica as mesmas. Para realizar esta atividade a NGA apoia o trabalho do revisor disponibilizando links para as bases de dados cadastradas de cada variante que o revisor está analisando.

Após o revisor finalizar a classificação o gestor é quem recebe esta lista, contendo os dados informados pelos revisores para então gerar o laudo.

As atividades do gestor e revisor na NGA pode ser visto na *Figura 15*.

Figura 15 – Atividades do gestor e revisor na NGA

Um usuário pode acumular dois ou até mesmo os três perfis existentes no NGA assim como um perfil pode ser composto por diversos usuários diferentes. Na *Figura 16* é possível observar o diagrama de caso de uso e verificar as atividades que cada perfil pode realizar.

Figura 16 – Diagrama de caso de uso de NGA

Na *Figura 17* é possível visualizar como é realizado o login no NGA.

Uma vez que um usuário pode pertencer a mais de um perfil, o próprio usuário no momento do *login* deve informar com qual perfil deseja entrar no sistema. Um botão para cada perfil é exibido na tela de *login* onde o usuário clica sobre o perfil desejado após informar seu e-mail e senha. Caso ele possua permissão de utilizar tal perfil o *login* é efetuado e as funções do perfil selecionado é disponibilizada.

Figura 17 - Login no NGA

Analise de Genoma

localhost:8080/analiseGenoma/

Index Register Log in

Log in

Email

Password

Manager Reviser Administrator Clear

5.1 Administrador

O administrador é o usuário responsável por gerenciar os usuários e bases de dados do sistema, podendo cadastrar as informações manualmente ou através da importação de arquivos tabular em um formato padronizado.

Na *Figura 18* é possível visualizar o *dashboard* do administrador.

Figura 18 - *Dashboard* do administrador

Patient Filter Analysis Options

Index

Select an option

5.1.1 Gerenciamento dos usuários

O cadastro de usuários é de acesso público, ao acessar a tela inicial o usuário pode ir à opção “*Register*” e realizar seu cadastro (*Figura 19*). Após o cadastro o usuário com acesso ao perfil administrativo deve liberar os acessos aos perfis que o usuário irá ter acesso.

Figura 19 - Cadastrado de usuário

The image shows a web browser window with the title 'Analise de Genoma'. The address bar displays 'localhost:8080/analiseGenoma/register'. The page content includes a navigation bar with 'Index', 'Register', and 'Log in' links. Below this is a section titled 'Create new account' containing five input fields: 'Name' (placeholder: 'Enter your name'), 'Email' (placeholder: 'Enter your email'), 'Password' (placeholder: 'Enter your password'), 'Replay password' (placeholder: 'Enter your password again'), and 'Phone' (placeholder: 'Enter your phonel'). At the bottom of this section are two buttons: 'Add' and 'Clear'.

Para liberar o acesso aos novos usuários, o usuário com perfil de administrador deve acessar a opção *User* no menu, uma lista com todos os usuários cadastrados será exibida (*Figura 20 - a*), ao selecionar um usuário será exibido o menu para modificar o usuário selecionado, neste o administrador poderá modificar os dados dos usuários, liberando ou bloqueando acesso aos perfis para os usuários cadastrado assim como modificar quaisquer informações dos usuários (*Figura 20 - b*).

Figura 20 - Gerenciar usuários

a)

Options	Id	Name	Email	Phone	Administrator	Manager	Reviser
Edit	2	Gestor da Silva	gestor@nsa.com	(016) 21019-300	true	true	false
Edit	1	Administrador da Silva	admin@nsa.com	(016) 21019-300	true	true	true
Edit	3	Marcelo Gomes de Paula	marcelogomesrp@	(016) 98245-0587	true	false	false
Edit	4	Mônica	monica@nsa.com.br	(016) 99999-9991	true	false	false
Edit	5	Cebolinha	cebolinha@nsa.com.br	(016) 99999-9992	true	false	false
Edit	6	Cascão	cascao@nsa.com.br	(016) 99999-9993	true	false	false
Edit	7	Magali	magali@nsa.com.br	(016) 99999-9994	true	false	false

b)

Users

Id: 4

Name: Mônica

Email: monica@nsa.com.br

Phone: (016) 99999-9991

Administrator: ☒

Manager: ☐

Reviser: ☐

Update Cancel

Options	Id	Name	Email	Phone	Administrat	Manager	Reviser
Edit	2	Gestor da Silva	gestor@nsa	(016) 21019-300	true	true	false
Edit	1	Administrato da Silva	admin@nsa	(016) 21019-300	true	true	true

Caso alguma informação não seja preenchida corretamente, o sistema alerta o usuário sobre o erro e destaca o campo com valor incorreto, um exemplo deste retorno ao usuário pode ser visto na *Figura 21*, onde tentou-se cadastrar um usuário sem informar valores nos campos obrigatórios.

Figura 21 - Alerta de erro

The screenshot shows a web browser window titled "Analise de Genoma" with the address bar displaying "localhost:8080/analiseGenoma/register.xhtml". The page has a navigation bar with "Index", "Register", and "Log in" links. Below this is a section titled "Create new account" with the following fields and error messages:

Field	Error Message
Name	Name must be longer than 3 characters
Email	Email isn't valid
Password	Password can't be null
Replay password	Password can't be null
Phone	Phone can't be null

The form also includes "Add" and "Clear" buttons at the bottom.

5.1.2 Gerenciamento das populações

Algumas mutações podem ser comuns para uma determinada população e não ser comuns para outra população, tornando estas informações importantes na seleção das variantes, por isso o sistema permite que no cadastro do paciente seja informado qual população ou subpopulação o paciente pertence.

As populações possíveis de serem selecionadas devem ser previamente cadastradas pelo administrador, este cadastro pode ser feito de forma manual, cadastrando um por um ou importando um arquivo *Comma-separated values* (CSV), separado por virgula as informações das populações que se deseja disponibilizar no sistema.

Na *Figura 22* é possível visualizar a tela do NGA que permite gerenciar as populações as populações cadastradas no NGA.

Figura 22 - Cadastro de população

User Population DB Bio DB Bio info Age Inheritance Gene Disease Options ▾

Etnia

Code*

Description

Super Population

New Find Cancel Import

Id	Code	Description	Super Population
1	AFR	African	
2	AMR	Ad Mixed American	
3	EAS	East Asian	
4	EUR	European	
5	SAS	South Asian	
6	CHB	Han Chinese in Beijing, China	EAS
7	JPT	Japanese in Tokyo, Japan	EAS
8	CHS	Southern Han Chinese	EAS
9	CDX	Chinese Dai in Xishuangbanna, China	EAS
10	KHV	Kinh in Ho Chi Minh City, Vietnam	EAS

A estrutura do arquivo a ser importado com os dados das populações pode ser visualizado na *Tabela 2*.

Tabela 2 - Descrição do arquivo de população

Coluna	Campo	Descrição
01	Código	Código de referência da população
02	Descrição	Descrição da população
03	Superpopulação	Caso a população a ser cadastrada seja uma subpopulação, deve ser informado o código da população.

5.1.2 Gerenciamento do banco de dados

Existem diversos banco de dados biológicos disponíveis atualmente. Manter todos estes bancos no sistema poderia torná-lo demasiadamente custoso, além de que as informações de muitos bancos de dados talvez nunca fossem utilizadas. Novos bancos de dados podem surgir ou mesmo atualizações dos existentes. Para não deixar o NGA demasiadamente pesado e

suportar outros ou mesmo novos bancos de dados biológicos, o NGA não fixa ou obriga a existência de nenhum banco de dados. Por este motivo o administrador deve cadastrar os bancos que o NGA irá trabalhar e importar seus dados.

Para cadastrar um novo banco de dados o administrador deve informar os seguintes dados: nome, URL e descrição onde apenas o nome é um campo obrigatório (*Figura 23*).

Por acreditar serem poucos os bancos de dados a ser criados de uma vez, não foi criada para esta atividade a opção de importação dos dados.

Figura 23 - Cadastro de banco de dados biológico

The screenshot shows a web application interface for managing biological databases. At the top, there's a navigation bar with tabs: 'User', 'Population', 'DB Bio', 'DB Bio info', 'Age', 'Inheritance', 'Gene', 'Disease', and 'Options'. The 'DB Bio' tab is selected. Below the navigation bar, there's a section titled 'Biological Databases'. It contains a form with three input fields: 'Name*' (placeholder: 'The database name'), 'URL' (placeholder: 'The database URL'), and 'Description' (placeholder: 'The database description'). Below the form is an 'Add' button. Underneath the form is a table with the following columns: 'Edit', 'Id', 'Name', 'URL', and 'Description'. The table contains three rows of data:

Edit	Id	Name	URL	Description
<button>Edit</button>	1	Orphanet	https://www.orpha.net/consor/cgi-bin/index.php	Banco de dados sobre doenças raras e medicamentos órfãos
<button>Edit</button>	3	OMIM	https://www.omim.org/	Um catálogo online de genes humanos e distúrbios genéticos
<button>Edit</button>	2	HPO	https://hpo.jax.org/app/	fornece um vocabulário padronizado de anormalidades fenotípicas encontradas em doenças humanas

O cadastro das informações do banco deve ser feito na opção que representa a informação. Por exemplo para cadastrar os genes existentes em um banco de dados deve ir à na opção gene e informar os dados do gene e o banco do que esta informação foi retirada. Estas informações dos bancos contém uma opção de importação.

5.1.3 Gerenciamento dos genes

Os genes são fundamentais para o correto funcionamento do NGA. Uma vez que um mesmo gene pode ter mais de um símbolo, o cadastro do gene, permite que se aponte seus sinônimos garantido o correto funcionamento do sistema, mesmo com divergência na

nomenclatura de símbolos por diferentes bancos de dados biológico.

O cadastro dos genes é realizado pelo administrador, que deve informar além dos dados referentes ao gene, o banco de dados de onde a informação foi obtida.

Os genes também são criados na importação do arquivo de variantes do paciente, neste caso se o gene da variante a ser importando não existir este é cadastrado automaticamente. É recomendado que o cadastro do gene seja realizado de forma antecipada, para garantir que seus sinônimos estejam cadastrados e evitar que o sistema mantenha informações de um gene, como se fossem dois genes diferentes.

Na *Figura 24* é possível visualizar a tela do NGA onde é realizada a manutenção dos genes, que pode ser realizada de forma manual, cadastrando gene a gene ou por meio da importação de um arquivo CSV, os campos deste arquivo podem ser visto na *Tabela 3*.

Figura 24 - Manutenção de genes

The screenshot shows a web browser window with the URL `localhost:8080/analiseGenoma/administrador/gene.shtml`. The page title is 'Analise de Genoma'. The main content area is titled 'Gene' and contains a form for adding or importing genes. The form has the following fields: 'Name*' (text input), 'Symbol*' (text input), 'Symbol Synonyms' (text input), 'Db bio*' (dropdown menu with 'select one' selected), 'Db Identifier' (text input), and 'URL' (text input). Below the form are three buttons: 'Add', 'Import', and 'Find'. Below the buttons is a table with the following columns: 'Edit', 'Id', 'Symbol', 'Name', 'Synonyms', 'Orphanet', 'OMIM', 'HPO', and 'HUGO'. The table contains two rows of data:

Edit	Id	Symbol	Name	Synonyms	Orphanet	OMIM	HPO	HUGO
Edit	1	BRCA1	BRCA1, DNA repair associated	BRCA1 RNF53 BRCA1/BRCA2-containing complex "BREAST CANCER 1 GENE; BRCA1"	15377	113705	672	1100
Edit	2	BRCA2	BRCA2, DNA repair associated	BRCA2 FANCD1 FANCD1 GENE; FANCD1	15378	600185	675	1101

Tabela 3 – Descrição do arquivo de genes

Coluna	Campo	Descrição
01	DB id	Id de identificação no banco de dados biológico
02	Símbolo	Símbolo do gene
03	Nome	Nome do gene
04	Sinônimos	Símbolo dos sinônimos
05	URL	URL com informações do gene

5.1.4 Gerenciamento das doenças

O cadastro das doenças é realizado pelo administrador, que deve informar os dados nome, CID, descrição, banco de dados biológico, idade, padrão de herança e prevalência, sendo apenas o nome e o banco de dados biológico são campos obrigatórios, os demais campos são opcionais.

O cadastro das informações pode ser realizado manualmente ou importando um arquivo CSV.

Na *Figura 25* é possível visualizar a tela que permite realizar a manutenção das doenças.

Figura 25 - Manutenção das doenças

Disease	Name	Icd	Description	Age	Inheritance	Prevalence
1	LONG CHAIN 3-HYDROXYACYL-COA DEHYDROGENASE DEFICIENCY	E71.3	LONG CHAIN 3-HYDROXYACYL-COA DEHYDROGENASE DEFICIENCY (LCHADD) IS A MITOCHONDRIAL DISORDER OF LONG CHAIN FATTY ACID OXIDATION CHARACTERIZED IN MOST PATIENTS BY ONSET IN INFANCY/EARLY CHILDHOOD WITH HYPOKETOTIC, HYPOGLYCEMIA, METABOLIC ACIDOSIS, LIVER DISEASE, HYPOTONIA AND FREQUENTLY CARDIAC INVOLVEMENT WITH ARRHYTHMIAS AND/OR CARDIOMYOPATHY.	Infancy, Neonatal	Autosomal recessive	9.0E-5
2	3-METHYLCROTONYL-COA CARBOXYLASE DEFICIENCY	E71.1	3-METHYLCROTONYL-COA CARBOXYLASE DEFICIENCY (3-MCCD) IS AN INHERITED DISORDER OF LEUCINE METABOLISM CHARACTERIZED BY A HIGHLY VARIABLE CLINICAL PICTURE RANGING FROM METABOLIC CRISIS IN INFANCY TO ASYMPTOMATIC ADULTS.	All ages	Autosomal recessive	9.0E-5
3	CRANIO-CEREBELLO-CARDIAC (3C) SYNDROME	E71.0	CRANIO-CEREBELLO-CARDIAC (3C) SYNDROME IS A RARE MULTIPLE CONGENITAL ANOMALIES SYNDROME CHARACTERIZED BY CRANIOFACIAL (PROMINENT OCCIPUT AND FOREHEAD, HYPERTELORISM, OCULAR PROXIMITY, DEEP-SEATED, CLEFT-LIP, CLEFT-PALATE)	Neonatal, Antenatal,	Autosomal recessive or	1.0E-5

Tabela 4 – Descrição do arquivo de doenças

Coluna	Campo	Descrição
01	Nome	Nome que identifica a doença
02	Descrição	Descrição da doença
03	ICD	Código do ICD
04	Hereditariedade	Padrão de hereditariedade
05	Idade	Idade que a doença se manifesta
06	Prevalência	Percentual de pessoas que possuem a doença

5.1.5 Gerenciamento das informações biológicas

As informações que ligam um determinado gene a uma doença também devem ser importadas pelo administrador onde esta informação deve estar associada a um banco de dados biológico.

Na *Figura 26* é exibido a tela que permite realizar a manutenção das informações biológicas.

O administrador pode cadastrar as informações biológicas manualmente, cadastrando uma por vez ou importando um arquivo CSV.

Para realizar a importação o administrador deverá informar o banco de dados a que se refere as informações que serão importadas (*Figura 27*). Os campos do arquivo CSV podem ser vistos na *Tabela 5*.

Figura 26 - Manutenção das informações biológicas

Analise de Genoma

localhost:8080/analiseGenoma/administrador/

User Population DB Bio DB Bio info Age Inheritance Gene Disease Options ▾

Db Bioinfo

Db Bio* select one ▾

Disease* Enter the disease

Genes* The genes symbo

URL Enter the URL that describe this diasease

Add Import List

Figura 27 – Importar informações biológicas

Tabela 5 - Descrição do arquivo de informações biológicas

Coluna	Campo	Descrição
01	Nome	Nome da doença
02	Gene	Gene relacionado a doença
03	Id	Identificador da informação no banco de dados biológico
04	URL	URL com as informações da doença no banco de dados

5.1.6 Opções

O administrador assim como todos os outros usuários, possui em seu menu a opção de trocar a senha, editar suas informações sair do sistema (*Figura 28 - Opções*).

Figura 28 - Opções

5.2 Gestor

O gestor é o usuário responsável por realizar as análises. Este usuário é quem deve cadastrar os pacientes, criar e gerar o laudo das análises.

5.2.1 Cadastro de paciente

Para cadastrar um paciente o gestor deve informar nome, data de nascimento, id secundário, população, gênero, pai, mãe e observação, onde os nome e data de nascimento são obrigatórios e os demais valores são opcionais (*Figura 29*).

Figura 29 - Cadastro de paciente

The screenshot shows a web browser window titled 'Analise de Genoma' with the URL 'localhost:8080/analiseGenoma/gestor/pacien'. The page displays a 'Patient Filter Analysis' section with a 'Patient' sub-section. The form contains the following fields:

- Name***: Text input with placeholder 'Enter the patient's name'.
- Birthday***: Text input with placeholder 'Enter the patient's birthday'.
- Second Id**: Text input with placeholder 'Enter the id used in your institution to identify this patient'.
- Population**: Dropdown menu with 'select one' and a downward arrow.
- Gender**: Dropdown menu with 'select one' and a downward arrow.
- Father**: Text input with placeholder 'Enter the name of the patient's father'.
- Mother**: Text input with placeholder 'Enter the name of the patient's mother'.
- Note**: Text area with placeholder 'use this area to write any important additional patient information'.

At the bottom of the form are two buttons: 'Save' and 'Cancel'.

O campo id secundário foi criado para que o gestor ao cadastrar um paciente, tenha a opção de informar o id gerado por outro sistema, facilitando assim a identificação do paciente.

Caso deseje realizar uma análise da família do paciente, os pais devem ser cadastrados primeiro e no cadastro do filho deve ser selecionado quem são os pais. Apenas para os pacientes que possuem pai ou mãe cadastrado terão a análise por família habilitada.

Um paciente pode ter um ou mais arquivos de variante cadastrado, este cadastro é realizado por meio de uma importação de um arquivo texto. A tela de *upload* do arquivo é exibida logo após o paciente ser adicionado no NSA (*Figura 30*).

Figura 30 – Importação do arquivo de variantes

The screenshot shows a web application interface for genomic analysis. A modal dialog titled 'Import variant file' is displayed in the center. The dialog contains the following fields and controls:

- Id:** 10
- Patient's name:** ALAN TURING
- age:** 0
- Variant file name:** Variant file name
- Data file:** A button labeled '+ Data file'.
- Buttons:** 'Add' and 'Cancel' buttons at the bottom.

The background shows a patient profile page with fields for Name, Birthday, Second Id, Population, Gender, Father, Mother, and Note. A 'Save' button is visible at the bottom left of the background page.

O formato de arquivo de entrada deve incluir as seguintes colunas: número de cromossomos, coordenada do cromossomo, alelo referência, alelo amostra, símbolo gene, transcrito ID, alteração de nucleotídeos, alteração de aminoácidos, SIFT por predição funcional, PolyPhen-2 predição funcional, phyloP p-value, dbSNP ID e 1000 frequência do alelo genoma. A descrição completa dos campos do arquivo de variantes pode ser vista no *Quadro 3*.

Quadro 3 - Descrição do arquivo de variantes

(Continua)

Coluna	Campo	Descrição
01	Cromossomo	Cromossomo que está a variante
02	Posição	Posição da variante
03	Variante	Nucleotídeos modificados
04	Símbolo do gene	Símbolo do gene da variante
05	Descrição gene	Descrição do gene da variante
06	UMD predictor	Valor atribuído pelo predictor UMD a variante

Quadro 3 - Descrição do arquivo de variantes

(Continuação)

Coluna	Campo	Descrição
07	Zygosity	Similaridade ou dissimilaridade do DNA entre cromossomos.
08	Allelic Depth	Fornece o número de variantes que suportam cada um dos alelos relatados.
09	Filter	Se a variante passou pelos controles de qualidade
10	HGVS_C	O nome da sequência codificadora do HGVS
11	HGVS_P	O nome da sequência proteica do HGVS
12	dbSNP id	Identificador da variante no DdSNP
13	Número do Exon/Intron	O número de exon/intron
14	Tipo	Tipos de genes e transcrição
15	Efeito	Termos de ontologia da variante
16	Impacto	Uma estimativa simples do impacto, que pode ser <i>High, Moderate, Low, Modifier</i>
17	Clinvar significance	Significado clínico da variante, podendo ser: <i>Uncertain significance, not provided, Benign, Likely benign, Likely pathogenic, Pathogenic, drug response, histocompatibility, other</i>
18	Clinvar disease name	Nome da doença associada a variante
19	Clinvar accession	Número de acesso e versão atribuído pela ClinVar à relação genótipo / fenótipo
20	Clinvar allele type	Um número que define os alelos nas colunas REF ou ALT. 0 é REF, 1 é o primeiro alelo ALT, etc. Um valor de -1 indica que nenhum alelo foi encontrado para corresponder a um nome de alelo HGVS correspondente.

Quadro 3 - Descrição do arquivo de variantes

(Continuação)

Coluna	Campo	Descrição
21	Clinvar allele origin	Descreve a origem do alelo da variante. Podendo ser: 0 - desconhecido; 1 - linha germinativa; 2 - somático; 4 - herdado; 8 - paterno; 16 - materno; 32 - de novo; 64 - biparental; 128 - uniparental; 256 - não testado; 512 - testado-inconclusivo; 1073741824 - outros
22	Sift	Predição Sift podendo ser Tolerated ou Damaging
23	Polyphen Hdiv	Predição do Polyphen2 baseado no hdiv
24	Polyphen Hvar	Predição do Polyphen2 baseado no hvar
25	Mutation taster	Predição do MutationTaster
26	LRT	Predição do annovar que pode ser <i>Deleterious</i> , <i>Neutral</i> ou <i>Unknown</i> .
27	GERP RS score	A pontuação para calcular a conservação de cada nucleótido no alinhamento de múltiplas espécies.
28	GERP neutral rate	Estimativas específicas de posição de restrição evolucionária usando estimativa de taxa evolutiva de máxima verossimilhança
29	Feature	Característica da variante, se é por exemplo uma transcrição
30	Ensembl ID	Identificação da variante no Ensembl
31	100 vertebrate genomes conservation score	Grau de conservação do sítio da mutação
32	Interpro Domain	Classificação no InterPro
33	Varint Status	Status da variante
34	Genotype	Representa a confiança em escala Phred que a atribuição do genótipo está correta
35	Read depth	Profundidade de leitura aproximada
36	Allele mut fraction	A fração de alelos na amostra que são mutações
37	Mean base quality	Pontuação do teste Z de a classificação de Wilcoxon da base alterada vs. qualidade base de referência

Quadro 3 - Descrição do arquivo de variantes

(Conclusão)

Coluna	Campo	Descrição
38	Variant type	Tipo de variante, utilizado no dbSNP
39	Validated	Validado por outro experimento
40	Donor splice site	Indica se a variante está dentro da região doadora de <i>splicing</i> do íntron
41	Acceptor splice site	Indica que a variante está localizada em um splice site
42	Is mutation	É mutação de baixa frequência que é citada em periódico e em outras fontes confiáveis
43	Frequência europeia	Frequência encontrada na variante na população europeia
44	Frequência africana	Frequência encontrada na variante na população africana
45	Frequência asiática	Frequência encontrada na variante na população asiática
46	Frequência americana	Frequência encontrada na variante na população americana
47	Frequência geral	Frequência encontrada na variante em todas as populações

A importação do arquivo de variante é realizada no cadastro do paciente pois esta é a operação mais demorada da fase de seleção de variantes, desta forma o gestor poderá criar diversas análises para um paciente sem perder o tempo de importação diversas vezes.

Caso deseje importar mais de um arquivo de variante é possível utilizar a opção “*add variant file*” disponível na busca de usuários.

5.2.2 Filtro

Ao criar uma análise, o gestor conta com um filtro padrão criado para selecionar as variantes de forma automática, porém este filtro pode não atender a todos os casos de análise por isso a ferramenta NGA permite ao gestor a criação de novos filtros utilizando os critérios de sua preferência.

Na *Figura 31* é possível visualizar parte da tela de criação de novos filtros da ferramenta NGA.

Figura 31 - Manutenção dos filtros

The screenshot displays a web browser window with two tabs: 'Análise de Genoma' and 'Google Tradutor'. The address bar shows 'localhost:8080/analiseGenoma/gestor/filter.xhtml'. The page title is 'Patient Filter Analysis' with a dropdown menu for 'Options'. The main content area is titled 'Filter' and contains a form with the following elements:

- A text input field labeled 'Name*' with the placeholder text 'Enter an name for this filter'.
- A checkbox labeled 'Select gene from analysis's disease'.
- A checkbox labeled 'Filter by Gene'.
- A checkbox labeled 'Filter by Chromosome'.
- A checkbox labeled 'Filter by Position'.
- A checkbox labeled 'Filter by Reference'.
- A checkbox labeled 'Filter by Changed'.

Com exceção do filtro “*Select gene from analysis's disease*” que irá utilizar a doença informada na criação da análise, os outros filtros requerem valores adicionais, por exemplo o filtro por gene, irá solicitar os símbolos dos genes que deseja selecionar e o filtro por posição irá solicitar o intervalo (*Figura 32*).

Para todos os filtros que possuem como entrada um valor uma cadeia de caracteres, a entrada dos dados é feita em um campo de busca que permite informar mais de um valor, como exibido na *Figura 32* para a busca de gene. Para valores numéricos são solicitados os valores mínimo e máximo como exibido na *Figura 32* para a busca por posição.

Figura 32 – Manutenção dos filtros, com gene e posição

The screenshot shows a web browser window with the URL `localhost:8080/analiseGenoma/gestor/filter.xhtml`. The page title is "Analise de Genoma". Below the browser window, there is a header bar with "Patient Filter Analysis" and "Options". The main content area is titled "Filter" and contains the following form elements:

- Name***: A text input field with the placeholder text "Enter an name for this filter".
- Select gene from analysis's disease**: A checkbox that is currently unchecked.
- Filter by Gene**: A checkbox that is currently checked.
- Select the genes**: A text input field.
- Filter by Chromosome**: A checkbox that is currently unchecked.
- Filter by Position**: A checkbox that is currently checked.
- Position**: A section containing two text input fields:
 - Min**: A text input field with the placeholder text "Enter the minimum value".
 - Max**: A text input field with the placeholder text "Enter the max value".

5.2.3 Análise

Para criar uma análise, o revisor deve informar um nome para a análise, o paciente e selecionar seu o arquivo de variantes, opcionalmente o gestor, pode também informar a doença suspeita, um ou mais pacientes correlatos e um filtro a ser aplicado.

O nome da análise é utilizado para identificar a mesma no sistema posteriormente.

A opção de aplicar o filtro padrão já vem pré-selecionada, porém o usuário tem a opção de desabilitar ou mesmo selecionar algum outro filtro.

Figura 33 - Nova análise

The screenshot shows a web browser window with the title 'Analise de Genoma'. The address bar shows 'localhost:8080/analiseGenoma/gestor/analise_nova.xhtml'. The main content area is titled 'Patient Filter Analysis - Options' and contains a 'New Analysis' section. The form has the following fields and controls:

- Name***: Text input with placeholder 'Name of new analysis'.
- Disease**: Text input with placeholder 'Disease name'.
- ICD**: Text input with placeholder 'Choose the disease to see the cid'.
- Note**: Text area with placeholder 'Some information that you need to remember to help you in this analysis'.
- Patient***: Text input with placeholder 'Patient's name'.
- Variant File ***: Dropdown menu with 'Choose one' selected.
- Correlate**: Text input with placeholder 'Correlate patient's name'.
- Variant file Correlate**: Dropdown menu with 'selecione one' selected.
- Add correlate**: Button next to the 'Variant file Correlate' dropdown.
- Correlates**: Section header above a table.
- Correlates selected**: Table header.
- No records found.**: Text inside the table.
- Apply pre-filter**: Radio button group with 'Yes' selected.
- Filter**: Dropdown menu with 'DEFAULT' selected.
- Add** and **Cancel**: Buttons at the bottom of the form.

5.2.4 Seleção de variantes

Ao criar uma análise e informar um filtro as variantes já são filtradas, selecionando apenas aquelas que atendem aos critérios do filtro. As variantes selecionadas são exibidas para que o gestor confirme a seleção (*Figura 34*). A ordem das colunas foi definida em reunião com o CMG de forma a facilitar este processo.

Figura 34 - Seleção de variantes

Analise de Genoma x

localhost:8080/analiseGenoma/gestor/analise_selecionar_variantes.xhtml

Patient Filter Analysis ▾ Options ▾

Select analysis variants

Analyze Patient Disease Filter

Id: 6
Name: Análise de pesquisa
Note:

Finalize selection

5 ▾ (1 of 1)

Id	Chr	Position.	ref	changed	Gene	Umd predictor	Zygosity
3063	2	8871342	C	A	KIDINS220	PROBABLY PATHOGENIC	HETEROZYGOUS
3277	19	50195648	T	C	CPT1C	PROBABLY PATHOGENIC	HETEROZYGOUS
1556	7	107593989	T	C	LAMB1	PATHOGENIC	HETEROZYGOUS

5 ▾ (1 of 1)

Caso os critérios sejam muito ou pouco restritivo o gestor tem a opção de personalizar os critérios do filtro da análise em andamento, clicando no cabeçalho de cada coluna o gestor poderá aplicar, remover ou modificar as opções de seleção daquele dado (Figura 35).

Figura 35 - Modificar opções do filtro

Todos os filtros aplicados podem ser consultados, clicando na aba *Filter*, onde será exibido os filtros aplicados, seus valores e a quantidade de total de variantes da análise e a quantidade selecionada após a aplicação dos filtros (Figura 36).

Figura 36 - Filtros aplicados

Patient Filter Analysis ▾ Options ▾

Select analysis variants

Analyze Patient Disease Filter

Filter name ANALISE DE PESQUISA Update filter name

full variant 360810

selected variant 8

Gene PAX3, GAN, SPG7, ALS2, RTN2, KLC2, ATL1, SPG21, PEX3, BSCL2,

UMD Predictor POLYMORPHISM, PATHOGENIC, PROBABLY PATHOGENIC,

Finalize selection

5 ▾ (1 of 2) >> >>

Id	Chr	Position.	ref	changed	Gene	Umd predictor	Zygoticity	Allelic depth
5407	2	223161889	A	G	PAX3	POLYMORPHISM	HOMOZYGOUS	0
5408	2	223161889	A	G	PAX3	POLYMORPHISM	HOMOZYGOUS	0
2579	14	51057727	G	A	ATL1	POLYMORPHISM	HETEROZYGOUS	127
2175	11	66031660	C	T	KLC2	POLYMORPHISM	HOMOZYGOUS	0
5131	2	202598113	C	T	ALS2	POLYMORPHISM	HETEROZYGOUS	153

5 ▾ (1 of 2) >> >>

Toda análise realizada criará um filtro com o mesmo nome da análise e este filtro pode ser utilizado como filtro para uma nova análise (Figura 36). Se o gestor desejar ele pode renomear este filtro de forma a facilitar sua localização e utilização em novas análises, este filtro personalizado pode ser escolhido no momento da criação de uma nova análise.

A aba *Disease* exibira as informações cadastradas referentes a doença da análise.

Após a seleção das variantes o gestor deve clicar no botão “*Finalize selection*”, que fará o gestor ir para a tela de seleção de revisores, estes revisores selecionados são os revisores que irão apoiá-lo na análise. As variantes selecionadas serão disponibilizadas aos revisores indicados para que possam classificar estas variantes (Figura 37).

Figura 37 - Seleção de revisores

The screenshot shows a web browser window titled 'Analise de Genoma'. The address bar displays 'localhost:8080/analiseGenoma/gestor/anal'. The main content area is titled 'Patient Filter Analysis - Options'. Below this, there is a section 'Select reviser'. Under the heading 'Analyse: Análise de pesquisa', there is a list of reviewers: 'Administrator da Silva' (unchecked) and 'Mônica' (checked). To the right of the list are four buttons: a right arrow, a right arrow with a plus sign, a left arrow, and a left arrow with a plus sign. At the bottom of the section is a 'Submit' button.

O gestor pode visualizar a situação da análise, para verificar o percentual de variante analisada para cada um dos revisores que a análise foi atribuída (*Figura 38*).

Figura 38 - Status da análise

The screenshot shows the 'Status' section of the 'Analise de Genoma' web application. It displays the following information:

- Name:** Análise de validação
- Reviser:** Administrator da Silva
- Progress:** 0.0%
- Reviser:** Mônica
- Progress:** 66.66666666666666%

5.2.5 Laudo da análise

Após a classificação dos gestores, o revisor deverá emitir um laudo da análise (*Figura 39*). Para isso o gestor terá acesso a todas as classificações e informações adicionais indicadas pelos revisores que é exibida ao clicar na aba *Revisers* (*Figura 40*) Nesta tela o gestor pode classificar individualmente cada variante e esta informação poderá ser utilizadas em outras análises. Caso o gestor não consiga concluir a análise com base nas informações dos revisores das variantes selecionadas, o gestor pode criar uma análise e utilizar outro critério para seleção de variantes.

Figura 39 – Finalizar laudo da análise

The screenshot displays the 'Analise de Genoma' web application interface. The browser window title is 'Analise de Genoma'. The address bar shows 'localhost:8080/analiseGenoma/gestor/analise_laudoar.xhtml'. The main content area is titled 'Patient Filter Analysis - Options' and contains a 'Laudar' section. This section includes a 'Report' header, a text area for 'Note', and a 'Patogenic' section with five radio button options: 'benign', 'likely benign', 'Uncertain significance', 'Likely pathogenic', and 'Pathogenic'. Below these options is a rich text editor toolbar with various formatting tools. At the bottom of the 'Laudar' section are 'Save' and 'finalize' buttons. A 'Revisers' section is visible at the very bottom of the page.

Figura 40 - Informações apoiar o laudo

Patient Filter Analysis - Options -

Laudar

Report

Revisers

Manager	Administrato da Silva	Mônica	Id	Chr	Position	ref	changed	Gene	Umd predictor
Opine	pathogenic	pathogenic	66515	19	50195648	T	C	CPT1C	PROBABLY PATHOGENIC
Opine	likely benign	likely benign	64794	7	107593989	T	C	LAMB1	PATHOGENIC
Opine	likely benign	benign	63541	2	8871342	C	A	KIDINS220	PROBABLY PATHOGENIC

O laudo será gerado em PDF, e o gestor tem a opção de realizar *download* deste arquivo e/ou imprimir para enviar aos interessados. Na Figura 41 é possível visualizar um exemplo de laudo.

Figura 41 - Exemplo de laudo

Patient Filter Analysis - Options -

Laudar

1 de 1

Zoom automático

NGA Report

Analysis: Análise de validação

Note:

Disease: SPASTIC PARAPLEGIA

Patient: MARIA

Second Id:

Birth: 1980-07-01

Gender: f

Lauded as: Pathogenic

Notes from manager

The Análise de validação analysis of patient MARIA was completed in Sat Jul 21 16:02:36 BRT

5.3 Revisor

O revisor é o especialista da doença em estudo que irá auxiliar o gestor no processo de análise. As análises indicadas pelo gestor são exibidas ao revisor que deverá então classificar cada uma das variantes da análise.

5.3.1 Classificação das variantes.

O revisor recebe a lista de variantes da análise do gestor e deve classificar cada variante de acordo com a sua patogenicidade. Para cada variante é exibido diversas informações (*Figura 42*).

Figura 42 - Visualização das variantes pelo revisor

answer	Id	Chr	Position	Ref	Changed	Gene	Umd predictor	Zygosity	Allelic deph	Filter	Hgvs_c
Opine	6651	19	50195648	T	C	CPT1C	PROBABLY PATHOGENIC	HETEROZYGOUS	55/15	PASS	C.139T>C
Opine	6479	7	107593989	T	C	LAMB1	PATHOGENIC	HETEROZYGOUS	12/6	PASS	C.3137A>G
Opine	6354	2	8871342	C	A	KIDINS220	PROBABLY PATHOGENIC	HETEROZYGOUS	120/122	PASS	C.4824G>T

Na classificação, o revisor consegue consultar de maneira rápida e fácil as informações da variante em questão. Na *Tabela 6* é possível visualizar as informações que possuem *links* e o banco de dados ao qual ela direciona. Além destes *links* também são adicionadas colunas para cada banco de dados biológico cadastrado, onde caso ele possua informações sobre esta variante/gene/doença será exibido um *link* que irá redirecionar diretamente para a página com as informações relacionadas.

Tabela 6 - Links de apoio ao revisor

Coluna	Banco	Site
Position	UCSC Genome Browser	https://genome.ucsc.edu/cgi-bin/hgGateway
Gene	Gene Cards	https://www.genecards.org
DbSNP	dbSNP	https://www.ncbi.nlm.nih.gov/projects/SNP/
Clinvar	Clinvar	https://www.ncbi.nlm.nih.gov/clinvar/
Ensembl	Ensembl	https://www.ensembl.org/Homo_sapiens/

O revisor deve classificar cada variante em: *benign*, *likely benign*, *uncertain significance*, *likely pathogenic* ou *pathogenic* e opcionalmente informar uma observação para justificar sua escolha ao gestor. Na Figura 43 é possível visualizar a tela que o gestor classifica a variante.

Figura 43 - Classificação da variante pelo gestor

The screenshot shows a web application interface for genomic analysis. A modal window titled "Opinar" (Opinion) is open, allowing a reviewer to classify a variant. The form includes the following elements:

- Header:** "Opinar" with a close button.
- Form Fields:**
 - Análise nome: Análise de validação
 - Variante: 6651913
 - Revisor: Mônica
 - Patogenic: Radio buttons for *benign*, *likely benign*, *uncertain significance*, *likely pathogenic*, and *pathogenic*.
- Rich Text Editor:** A toolbar with options for font (Sans Serif), size (Normal), bold (B), italic (I), underline (U), strikethrough (ABC), text color (A), background color (A), subscript (x₂), superscript (x²), heading (H₁, H₂), quote, and code. Below the toolbar is a large text area for notes.
- Action:** An "Aplicar" (Apply) button at the bottom left.

5.4 Arquitetura

Para facilitar o acesso dos usuários a ferramenta foi desenvolvida para a web. Desta forma, além de não depender de instalações no cliente, a NGA pode ser acessado por qualquer dispositivo que possua um navegador compatível.

O sistema foi construído em Java utilizando o SGBD Postgresql, utilizando os frameworks JSF, JPA e Primefaces. Para facilitar o desenvolvimento/instalação o mesmo foi criado utilizando um projeto do tipo Maven.

O sistema desde seu início foi desenvolvido e armazenado no Github, que proporcionou além de um controle de versão, um *backup* das fontes uma vez que o sistema coexistia nas máquinas utilizadas no desenvolvimento e no servidor.

O sistema segue o padrão MVC implementado no JSF, onde o *Model* foi dividido nas camadas de persistência e serviço.

A camada de persistência é a camada lógica que faz o acesso ao SGBD e a de serviço é a que contém as validações e a lógica da aplicação.

A camada *View* foi dividida em 3 camadas lógica, separando as *views* de acordo com o perfil do usuário (gestor, revisor e administrador).

A camada de *Controller* é implementada pelo JSF.

5.5 Implementação do sistema

O sistema foi desenvolvido em Java 8, utilizando o framework JavaServer Faces (JSF) na versão 2, que permite a construção de interfaces de usuário baseada em componentes para aplicação WEB. O Primefaces na versão 6.1 foi utilizado junto com o JSF, o que permitiu a criação de interfaces Ajax de forma simples.

O acesso ao banco de dados é realizado utilizando o Java Persistence API (JPA) o que garante uma independência do SGBD. No desenvolvimento foi utilizado o SGBD PostgreSQL na versão 10.

Para gerenciar as dependências o Maven foi utilizado. O sistema foi construído com 3 camadas físicas: camada de apresentação, camada de aplicação e camada de dados (*Figura 44*).

Figura 44 - Camadas físicas do NGA

A estrutura lógica do sistema foi organizada em várias camadas, como mostrado na Figura 45: uma camada para armazenar a *views* (Páginas Web) que se comunica com os *backbeans* armazenados nas camadas *managedbean* e *sessionbean*. Os *backbeans* estão separados de acordo com o seu tempo de vida, onde os *managedbean* são válidos enquanto a página está ativa e no *sessionbean* os objetos são ativos durante toda a sessão, tanto os *sessionbeans*, como os *managedbeans* apenas fazem chamadas aos métodos existentes na camada de serviço, que é a camada que possui todos as regras de negócio do NGA. O acesso ao banco de dados ficou isolada na camada chamada DAO (*data access object*) e o todos os modelos da aplicação são mantidos na camada chamada de *model*. Outras camadas lógicas foram criadas para abrir códigos auxiliares, como por exemplo a *convert* que armazenam conversores.

A importação dos dados do paciente no sistema não se mostrou eficiente utilizando o JPA, uma vez que para cada linha/campo do arquivo é preciso fazer uma busca e caso o valor procurado não seja encontrado, é preciso inserir este valor, antes de inserir o campo da linha. Por exemplo, para inserir uma variante que está no gene BRCA1 é necessário buscar por este gene antes. Caso o gene não tenha sido cadastrado ele deve ser cadastrado para que a variante seja cadastrada. Esta sequência de busca e inserção é realizada para diversos campos da tabela.

Figura 45 - Camadas lógicas do NGA

Para melhorar o tempo de importação, foi criada uma *Stored Procedure* (SP) que recebe os dados de uma variante e realiza todo o processamento necessário para cadastrar esta variante. A chamada a esta SP é realizada utilizando JSR 352 (processamento batch). Uma outra SP foi criada para gerar os metadados das variantes (quantidade de variantes, genes presentes, valores máximos e mínimos para posição etc.). Estes dados são utilizados como critério nos filtros e são exibidos para o gestor para apoiá-lo na análise.

A NGA está disponível para *download* no site <https://github.com/marcelogomesrp/analiseGenoma> para a comunidade científica, e interessados possam utilizar e modificar quando julgarem necessário.

6 VALIDAÇÃO

Nesse capítulo estão descritos os processos utilizados para validação do NGA.

A fim de preservar o anonimato, os nomes dos pacientes e o nome dos pesquisadores que responderam à pesquisa foram alterados ou omitidos.

A validação foi realizada de duas formas: validação da usabilidade e um estudo de caso.

A validação da usabilidade foi realizada aplicando um questionário a pesquisadores e profissionais de saúde.

A validação prática, foi realizada refazendo uma análise cujo laudo é conhecido a fim de garantir que o sistema está selecionado corretamente as variantes e avaliar suas vantagens e possivelmente desvantagens.

6.1 Avaliação da usabilidade

Para avaliar a usabilidade do NGA, foi aplicado um questionário contendo perguntas sobre a usabilidade geral do sistema. As perguntas têm o intuito de averiguar se o sistema é amigável e se atende às necessidades de seus possíveis usuários.

Para tanto, foram convidados 11 potenciais usuários do sistema que utilizaram o NGA em uma análise conduzida, onde enquanto o usuário recebia orientações dos passos necessários para realizar uma análise ele utilizava o sistema para realizar uma análise. Posteriormente o questionário de usabilidade geral do sistema foi aplicado.

Foram convidados a participar desta avaliação pessoas que utilizam sistemas computacionais e que realizam ou que estão envolvidos em projetos que será necessário a análise de NGS sendo que das doze pessoas que responderam as questões oito afirma já ter realizado análise de exoma em projeto de pesquisa, das doze pessoas seis também utilizam análise de NGS para diagnóstico e entre todas as doze pessoas três realizam análise de exoma rotineiramente.

As questões aplicadas para avaliar a usabilidade do NGA podem ser visualizadas na *Tabela 7*.

Tabela 7 - Questões aplicadas na avaliação de usabilidade

Id	Descrição
Q01	A facilidade no aprendizado do uso do software
Q02	A facilidade na identificação das funções
Q03	A facilidade na identificação do que deve ser informado em cada campo
Q04	Sou capaz de realizar o meu trabalho de forma eficaz usando este sistema
Q05	A utilização do sistema apresenta ganhos com relação a metodologia utilizada atualmente
Q06	Acredito que a análise será mais produtiva com o uso deste sistema
Q07	Acredito que a análise terá resultado mais confiável com o uso deste sistema
Q08	Me senti confortável usando este sistema
Q09	Quando cometi algum erro no sistema é fácil identificar e resolver o erro
Q10	As informações visuais que o sistema oferece são fáceis de entender
Q11	As informações que o sistema oferece são eficazes em me ajudar a completar as tarefas
Q12	A organização de interface do sistema é clara
Q13	Minha satisfação com o sistema de forma geral

Estas questões foram respondidas com valores no intervalo de 1 até 7, sendo que o número 1 significa que o usuário discorda fortemente enquanto 7 significa que o usuário concorda fortemente. A opção “não se aplica”, foi utilizada quando o usuário não sabe opinar a respeito da pergunta. Na *Tabela 8* é exibido resultado das questões para cada usuário e a média de pontuação obtida em cada questão.

Tabela 8 - Respostas de cada usuário para o questionário de usabilidade

Questão	Usuários											Média
	01	02	03	04	05	06	07	08	09	10	11	
Q01	6	6	5	6	5	7	6	7	6	7	7	6,18
Q02	6	6	6	6	6	7	7	7	6	7	7	6,45
Q03	5	7	6	6	6	6	7	7	6	7	7	6,36
Q04	6	5	7	6	5	6	6	6	6	7	7	6,09
Q05	7	5	7	7	7	7	NA	7	6	6	6	6,50
Q06	7	7	7	6	7	7	6	7	7	7	7	6,73
Q07	7	5	7	6	7	6	NA	5	7	7	7	6,40
Q08	6	7	6	7	6	6	7	7	6	6	6	6,45
Q09	5	7	7	6	7	6	7	7	6	6	7	6,55
Q10	5	6	6	5	5	6	7	7	7	7	7	6,09
Q11	6	6	7	5	6	6	7	7	6	6	6	6,18
Q12	5	5	5	6	5	6	6	6	6	6	6	5,73
Q13	5	6	6	7	6	6	6	7	6	6	7	6,27

Sobre a facilidade no uso do *software*, que é a questão 1 do questionário pode ser observado que a nota média é de 6,18 tendo mais de 35% das pessoas classificando com a nota máxima e nenhuma nota inferior a cinco. Podemos concluir que a funcionalidade do *software* ser de fácil uso foi atendida.

As funções do sistema são apresentadas em um menu, exibindo todas as funções disponíveis para o usuário com o perfil selecionado. Ao omitir as funções de outros perfis o que torna o menu mais limpo e facilita ao usuário identificar as funções disponíveis para ele.

O que pode ser comprovado com a avaliação realizada na questão 2, onde obteve um média de 6,5 tendo 50% dos usuários classificando com nota máxima.

O uso de legendas claras e *placeholder* nos campos facilitou a identificação do valor que deve ser informado em cada campo, o que pode ser comprovado com a questão 3, onde 50% das pessoas responderam com a nota máxima a esta avaliação.

Os filtros, *links* e fluxo de trabalho da NGA permite a realização da análise de exoma de forma eficiente, tendo uma média acima de seis (6,09) na avaliação.

Quando comparado o uso do NGA com relação a metodologia utilizada atualmente 60% das pessoas que responderam ao questionário classificaram com sete que é a nota máxima permitida, tendo para esta questão uma média de 6,5.

A NGA possui o filtro padrão que automatiza a parte mais trabalhosa da análise, além

de possuir um visualizador que permite acessar de forma fácil informações externas. Estas funcionalidades foram adicionadas com o intuito de deixar a ferramenta mais produtiva o que é confirmado na pesquisa que teve esta questão classificada como 7 (nota máxima) por 72,7% das pessoas que responderam ao questionário.

A análise de exoma, tem uma dependência muito grande do conhecimento da pessoa que está conduzindo a mesma, uma vez que a ferramenta auxilia nesta atividade reduzindo esta dependência, torna o resultado da análise mais confiável.

A NGA utiliza componentes do Primefaces que é utilizado em diversas ferramentas web, o que permite ao usuário já ter uma familiaridade com a ferramenta mesmo sem nunca ter utilizado, isso é comprovando com a pesquisa, onde na questão se o sistema deixa o usuário confortável em seu uso, obteve apenas notas 7 e 6 (as duas maiores possíveis) para esta questão.

Ao ter algum preenchimento inválido o sistema alerta com uma mensagem e destacam o campo inválido para facilitar a identificação e correção do erro, como exibido na *Figura 21*, esta característica agradou aos utilizadores que classificou com a média de 6,55.

As informações visuais que o sistema oferece segue padrões, por exemplo todos os campos têm seu valor indicado por um texto a esquerda, campos com erros são pintados de vermelho, todos os alertas aos usuários são exibidos no topo e a direita página. A questão da pesquisa, que verifica se as informações são fáceis de entender, foi avaliada com notas seis e sete por 72,8% das pessoas.

As informações são agrupadas e exibidas em abas por exemplo, isso permite aos usuários não perder tempo interpretando as informações que são exibidas pelo sistema. As informações que o sistema oferece são eficazes em ajudar a completar as tarefas.

Foram criados três perfis de usuário, para evitar disponibilizar funções que nunca serão utilizadas por determinados usuários e mesmo que um usuário tenha acesso a todos os perfis, ele consegue se entrar com um ou outro e visualizar apenas os recursos inerentes ao perfil selecionado, facilitando a organização de interface do sistema. O NGA atendeu a este objetivo, pois foi avaliado na pesquisa com uma média de 5,73.

A satisfação geral com o sistema obteve uma média de 6,27 o que indica que o sistema atende as necessidades dos usuários.

Dadas as respostas obtidas pela aplicação do questionário, pode ser constatado que o sistema apresenta uma usabilidade satisfatória para a realização de suas tarefas.

6.2 Estudo de caso

A fim de avaliar a ferramenta do ponto de vista prático, uma análise foi realizada utilizando a NGA por um especialista que atua com bioinformática no HCRP-USP em projetos de bioinformática incluindo testes genéticos com dados de NGS e teve o apoio de uma aluna de doutorado em genética cujo estudos são na área de análise de exoma, que exerceu o papel de revisor neste estudo de caso.

Os dados disponibilizados do paciente do sexo feminino e a doença suspeita é *Spastic Paraplegia* (Paraplegia espástica). O arquivo de entrada possui 360.810 variantes, com 51.693 genes diferentes.

O paciente foi cadastrado com o nome fictício de “Maria” e a análise foi criada com o nome de “Análise de validação”, utilizando o filtro padrão (*Figura 46*).

Figura 46 – Criação da análise de validação

The screenshot shows a web browser window with the title 'Analise de Genoma'. The address bar shows 'localhost:8080/analiseGenoma/gestor/analise_nova.xhtml'. The main content area is titled 'Patient Filter Analysis - Options' and contains a 'New Analysis' form. The form has the following fields and values:

- Name***: Análise de validação
- Disease**: SPASTIC PARAPLEGIA
- ICD**: Choose the disease to see the cid
- Note**: Some information that you need to remember to help you in this analysis
- Patient***: MARIA
- Variant File ***: arquivo completo
- Correlate**: Correlate patient's name
- Variant file Correlate**: seleccione one
- Correlates**: Correlates selected (No records found.)
- Apply pre-filter**: Yes (checked)
- Filter**: DEFAULT

At the bottom of the form are 'Add' and 'Cancel' buttons.

A aplicação do filtro padrão, selecionou 3 variantes (rs20556, rs140961441, rs1044280) de um total de 360.810, eliminando mais de 99% das variantes a serem revisadas (*Figura 47*).

Figura 47 - Seleção das variantes na análise de validação

Umd predictor	Zygosity	Allelic depth1	Allelic depth2	Filter	Hgvs_c	Hgvs_p	Dbsnp	Exon/Intron number	
Pathogenic	HETEROZYGOUS	12	6	PASS	C.3137A>G	p.Gln1046Arg	rs20556	20	PROTEIN
Probably pathogenic	HETEROZYGOUS	55	15	PASS	C.139T>C	p.Trp47Arg	rs140961441	2	PROTEIN
Probably pathogenic	HETEROZYGOUS	120	122	PASS	C.4824G>T	p.Gln1608His	rs1044280	30	PROTEIN

A NGA se mostrou eficaz na seleção das variantes, uma vez que conseguiu eliminar uma grande quantidade de variantes a ser analisada, facilitando o trabalho do revisor que terá menos variantes para classificar.

As variantes então foram submetidas ao revisor, o qual, classificou as variantes para efeito de pesquisa e de "vigilância" identificadas no DbSNP com os Ids rs1044280 e rs140961441 como provavelmente patogênicas.

O revisor justificou a classificação pelo fato de que estas mutações estão classificadas como patogênicas ou provavelmente patogênicas em duas ferramentas de predição, UMD e na ferramenta CADD - que compila Sift, Polyphen, informações de proteína, mRNA e etc.

Com base nas informações do revisor, o gestor verificou que a variante rs140961441 além de estar classificada como "Disease causing" pelo mutation taster, também tem um valor mais alto no gerp rs score e também no vertebrate genomes o que pode ser um indicativo que esta variante pode estar em uma região importante o que mutação nesta variante pode ocasionar a doença.

Com base nas informações do revisor e suas investigações o gestor pode então emitir o laudo, apontando a variante rs140961441 como provavelmente patogênica.

Este caso havia sido anteriormente analisado de forma manual por um especialista e a variante classificada foi a mesma que a da NGA, confirmado, dessa forma, a eficácia da ferramenta.

Neste caso o filtro padrão foi utilizado e se mostrou eficiente. Em outros casos, dependendo das especificidades, o gestor tem na NGA a possibilidade de modificar os parâmetros aplicados sempre que julgar necessário, tornando o filtro mais ou menos restritivo.

6.3 Considerações finais

Para efeito de comparação dos resultados, tentou-se realizar uma análise com as ferramentas correlatas sem sucesso.

O capítulo seguinte apresenta uma discussão um pouco mais aprofunda dos resultados obtidos e fará uma síntese do trabalho realizado até aqui. Ainda, serão propostos trabalhos futuros a serem realizados com o NGA.

7 CONCLUSÃO

Ao longo desse trabalho foram descritos o modelo proposto, e a implementação da ferramenta NGA assim como a sua validação. Procurou-se, em cada um desses momentos, ressaltar os problemas enfrentados e justificar as decisões tomadas no seu desenvolvimento. Nesse capítulo serão discutidos os resultados apresentados no capítulo anterior e serão projetados trabalhos futuros no intuito de aperfeiçoar a ferramenta NGA.

A modelagem do processo de análise foi realizada a seguindo os requisitos levantados junto aos especialistas. A ferramenta NGA foi desenvolvida seguindo o modelo proposto que permite além de apoiar a análise de NGS registrar o conhecimento da relação entre a doença suspeita de diagnóstico e o gene. Este conhecimento é obtido nas análises conduzidas na NGA e pode ser utilizado em novas análises.

A usabilidade e funcionalidades da ferramenta foram avaliadas conforme descrito no capítulo 5 e ela obteve uma média satisfatória para usabilidade. As funcionalidades estão em conformidade, sendo assim foi possível realizar a análise descrita no capítulo 5 em um tempo menor e com menor chance de erros humanos quando comparada à análise realizada sem o uso da NGA.

O estudo de caso descrito na seção 6.2 apresenta a análise de uma paciente, que foi feita utilizando a NGA. Ela foi conduzida por um bioinformata, que teve apoio de uma aluna de doutorado em genética cujo estudos são na área de análise de exoma, que exerceu o papel de revisor neste estudo de caso.

Este caso havia sido anteriormente analisado com o uso de planilha eletrônica por outro profissional e a variante classificada nos dois casos é a mesma, confirmando a eficácia da ferramenta.

O resultado nos dois casos foram os mesmos, porém a NGA neste estudo de caso, possibilitou chegar ao resultado mais rápido e com maior comodidade aos usuários.

Tentou-se realizar o estudo de caso descrito na seção 6.2 utilizando as ferramentas correlatas, porém por falha no *software* ou mesmo de documentação, não se obteve sucesso nas análises.

Como informado anteriormente, a ferramenta fica disponível para *download* no *site* <https://github.com/marcelogomesrp/analiseGenoma> para a comunidade científica, e interessados possam utilizar e modificar quando julgar necessário.

Trabalhos futuros

As principais melhorias no modelo e na ferramenta que se sugerem como trabalhos futuros são: 1) Incorporara o *pipeline* de pré-processamento à ferramenta; 2) utilizando o conhecimento gerado na NGA para treinamento, criar uma ferramenta com inteligência artificial para apoiar o gestor.

REFERÊNCIAS

ANNILO, T e colab. **The human ribosomal protein S7-encoding gene: isolation, structure and localization in 2p25**. *Gene*, v. 165, n. 2, p. 297–302, 20 Nov 1995. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/8522193>>. Acesso em: 27 jul 2018.

AUTON, Adam e colab. **A global reference for human genetic variation**. *Nature*, v. 526, n. 7571, p. 68–74, 30 Set 2015. Disponível em: <<http://www.nature.com/doifinder/10.1038/nature15393>>. Acesso em: 25 jul 2016.

CARVALHO, Mayra Costa da Cruz Gallo De e SILVA, Danielle Cristina Gregorio Da. **Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas**. *Ciência Rural*, v. 40, n. 3, p. 735–744, Mar 2010. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-84782010000300040&lng=en&nrm=iso&tlng=pt>. Acesso em: 12 set 2015.

ELMASRI, Ramaze. **Fundamentals of database systems**. 6. ed. Boston: Addison-Wesley, 2010.

FARAH, Solnge Bento. **DNA: segredos & mistérios**. 2. ed. São Paulo: Sarvier Editora de livros médicos Ltda, 2007.

GUIMARÃES, Raony e DO, Corrêa e LISBÔA, Carmo. **Mendel , MD : um software online para análise de exomas humanos e investigação de doenças**. 2015.

ILLUMINA. **Illumina VariantStudio Data Analysis Software**. p. 1–4, 2016.

JOHNSON, Alberts e colab. **Biologia Molecular da Célula**. [S.l: s.n.], 2017. v. 6º edição.

KAMPS, Rick e colab. **Next-Generation Sequencing in Oncology : Genetic Diagnosis , Risk Prediction and Cancer Classification**. 2017.

KLUG, William S. e colab. **Conceitos de Genética**. 9. ed. Porto Alegre: Artmed, 2010.

KÖHLER, Sebastian e colab. **The Human Phenotype Ontology in 2017**. Nucleic Acids Research, v. 45, n. D1, p. D865–D876, 4 Jan 2017. Disponível em: <<https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1039>>. Acesso em: 15 jun 2018.

MCKUSICK, Victor A. **Mendelian Inheritance in Man and its online version, OMIM**. American journal of human genetics, v. 80, n. 4, p. 588–604, Abr 2007. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/17357067>>. Acesso em: 11 jun 2016.

MCLAREN, William e colab. **The Ensembl Variant Effect Predictor**. Genome Biology, v. 17, n. 1, p. 122, 6 Dez 2016. Disponível em: <<http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4>>. Acesso em: 24 jun 2018.

Orphanet. Disponível em: <<https://www.orpha.net/consor/cgi-bin/index.php>>. Acesso em: 15 jun 2018.

PARK, Young M e colab. **The EBI search engine: EBI search as a service-making biological data accessible for all**. Nucleic acids research, v. 45, n. W1, p. W545–W549, 3 Jul 2017. Disponível em: <<https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx359>>. Acesso em: 28 jul 2018.

REES, David C. e WILLIAMS, Thomas N. e GLADWIN, Mark T. **Sickle-cell disease**. The Lancet, v. 376, n. 9757, p. 2018–2031, 2010. Disponível em: <[http://dx.doi.org/10.1016/S0140-6736\(10\)61029-X](http://dx.doi.org/10.1016/S0140-6736(10)61029-X)>.

SALGADO, David e colab. **UMD-Predictor: A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution**. Human Mutation, v. 37, n. 5, p. 439–446, Maio 2016. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/26842889>>. Acesso em: 24 jun 2018.

SAM, The e FORMAT, B A M e WORKING, Specification. **Sequence Alignment/Map Format Specification**. n. May, p. 1–20, 2018. Disponível em: <<https://samtools.github.io/hts->

specs/SAMv1.pdf>.

SAMTOOLS. The Variant Call Format Specification v4.3. Online Resource, p. 1–28, 2015.

Disponível em: <<https://samtools.github.io/hts-specs/VCFv4.3.pdf>>.

SHORTLIFFE, Edward H. e CIMINO, James J. Biomedical Informatics. [S.l: s.n.], 2015.

WANG, Jinlian e colab. ClinLabGeneticist: a tool for clinical management of genetic variants from whole exome sequencing in clinical genetic laboratories. Genome Medicine, v. 7, n. 1, p. 77, 2015.

Disponível em: <<http://genomemedicine.com/content/7/1/77>>.

WHO | ICD-11 Revision. WHO, 2018. Disponível em: <<http://www.who.int/classifications/icd/revision/en/>>. Acesso em: 24 jun 2018.

YATES, Bethan e colab. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic acids research, v. 45, n. D1, p. D619–D625, 4 Jan 2017. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/27799471>>. Acesso em: 28 jul 2018.

APÊNDICE A

Estatísticas geradas pela aplicação do questionário de usabilidade.

A utilização do sistema apresenta ganhos com relação a metodologia utilizada atualmente

Acredito que a análise será mais produtiva com o uso deste sistema

Acredito que a análise terá resultado mais confiáveis com o uso deste sistema

Me senti confortável usando este sistema

Quando cometi algum erro no sistema é fácil identificar e resolver o erro

As informações visuais que o sistema oferece são fáceis de entender

**As informações que o sistema oferece
são eficazes em me ajudar a completar as tarefas**

A organização de interface do sistema é clara

Minha satisfação com o sistema de forma geral