UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

PEDRO BESCHORNER MARIN

Exomim: uma aplicação *web* para priorização de variantes em *whole-exome sequencing*

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Claudio Fernando Resin Geyer

Porto Alegre
2015

“We are glorious accidents of an unpredictable process with no drive to complexity, not the expected results of evolutionary principles that yearn to produce a creature capable of understanding the mode of its own necessary construction.”

— Stephen Jay Gould

AGRADECIMENTOS

Gostaria de agradecer a todas as pessoas que, de alguma forma, estiveram envolvidas em meu processo de graduação. À Universidade Federal do Rio Grande do Sul e ao meu orientador pela oportunidade. Aos pesquisadores do Hospital de Clínicas de Porto Alegre e do Instituto de Informática pela parceria na realização desse trabalho. Aos meus colegas e amigos que foram extremamente importantes durante esse período. A minha família e namorada pelo amor e apoio incondícionais durante todos esses anos da minha vida. A todos que aqui estão ou que já se foram, obrigado!

RESUMO

Para a identificação de mutações de DNA que possam prejudicar a saúde de seus portadores, a genética médica concentra esforços na análise de um importante trecho do genoma humano, o exoma. O grande volume de dados provenientes do sequenciamento genético e das inúmeras bases de dados médicos disponíveis na *web* faz com que ferramentas de Bioinformática sejam indispensáveis nesse processo. Uma etapa decisiva na análise de exomas é a de priorização das variantes genéticas. Atualmente, existem diversas ferramentas de priorização de variantes disponíveis. Entretanto, encontrar uma adequada às necessidades de um grande centro de pesquisas é uma tarefa difícil. Por isso, foi desenvolvido o Exomim, uma aplicação *web* para priorização de variantes que utiliza dados fenotípicos como elemento extra de filtragem. Os fenótipos são utilizados para a montagem de uma lista de genes correlacionados. Esses dados são provenientes do *Online Mendelian Inheritance in Man*, uma conceituada fonte de informação médica disponível na Internet. Nesse trabalho foi realizado o projeto, a implementação de um protótipo e a análise dessa ferramenta. Como resultado, obteve-se uma aplicação capaz de filtrar até 14 campos diferentes de variantes anotadas. Um desses campos está diretamente relacionado aos genes buscados a partir de fenótipos. Espera-se que esse trabalho possa contribuir para futuras pesquisas na área das análises clínicas.

Palavras-chave: Exoma. priorização de variantes. *web*. OMIM.

Exomim: a web application for variant prioritization in whole-exome sequencing

ABSTRACT

To identify DNA's mutations that may endanger the health of their human carriers, the medical genetics concentrates its efforts on the analysis of an important section of the human genome, the exome. The large volume of data from genome sequencing and the numerous medical consultation databases available in the web, makes Bioinformatics tools indispensable in this process. A turning point in exomas analysis is in the genetic variants prioritization. Currently, there are several tools available for variants prioritization. However, finding an adequate to the needs of a major research center is a difficult task. Therefore, we developed the Exomim, a web based application to prioritize variants using phenotypic data as an extra filter element. The phenotypes are used for assembling a list of their related genes. These data are sourced from the Online Mendelian Inheritance in Man, a respected medical information database available on the Internet. In this work it was carried out the project, the implementation of a prototype and the analysis of this tool. As a result, we obtained an application capable of filtering up to 14 different fields of annotated variants. One of these fields is directly related to genes fetched from phenotypes. It is to be hoped that this work will contribute to future research in the area of clinical analysis.

Keywords: exome, variant priorization, web, OMIM.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

NGS	Next-Generation Sequencing
WES	Whole-Exome Sequencing
CCSD	Consensus Coding Sequence Database
WGS	Whole-Genome Sequencing
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
NHGRI	National Human Genome Research Institute
MeSH	Medical Subject Headings
NLM	National Library of Medicine
OMIM	Online Mendelian Inheritance in Man
VCF	Variant Call Format
VEP	Variant Effect Preditor
DSF	Django Software Foundation
MTV	Model-Template-View
MVC	Model-View-Controller
HTML	HyperText Markup Language
HCPA	Hospital de Clínicas de Porto Alegre
SQL	Structured Query Language
JHUSOM	Johns Hopkins University School of Medicine
ASCII	American Standard Code for Information Interchange
API	Application Programming Interface

SUMÁRIO

1 INTRODUÇÃO

O *whole-exome sequencing* (WES) é uma das aplicações para sequenciamento de material genético mais utilizadas atualmente (CHOI et al., 2009). O resultado dessa técnica fornece a pesquisadores e profissionais da saúde um conjunto de milhares de variantes genéticas pertencentes a uma parte importante do genoma do paciente, o exoma. Esse segmento do genoma é de suma relevância para análises clínicas, pois é a região que codifica as proteínas que estarão presentes nas mais diversas vias metabólicas do organismo (GRIFFITHS et al., 2012).

Denomina-se variante genética qualquer diferença que possa ocorrer entre nucleotídeos em uma mesma posição do DNA para indivíduos de mesma espécie. A ocorrência dessa variação é natural e, na maioria das vezes, não favorece nem prejudica o seu portador. Entretanto, existem trechos do genoma humano onde as consequências dessas modificações são mais severas e afetam diretamente a saúde de seu portador sendo o exoma um deles. Das doenças genéticas cuja base molecular é conhecida, mais de 80% delas estão relacionadas a variantes localizadas no conjunto de éxons do DNA (CHOI et al., 2009).

Uma das etapas do fluxo de trabalho em um WES é a priorização do conjunto total de variantes identificadas, cujo objetivo é selecionar aquelas potencialmente prejudiciais à saúde do paciente, e que possam contribuir para o direcionamento do diagnóstico clínico por parte da equipe médica (MANOLIO et al., 2009; BAMSHAD et al., 2011; KIEZUN et al., 2012). A análise de priorização de exomas é realizada por diferentes métodos heurísticos. O método mais utilizado é caracterizado por filtragens sucessivas e é consenso que se utilize pressupostos principais para definir uma única variante causal dentre as milhares presentes. De uma forma geral, os pressupostos assumem que: (1) a variante causal irá alterar a sequência codificante da proteína; (2) será extremamente rara; (3) a penetrância será completa; e (4) todo indivíduo com determinada desordem irá apresentar a suposta variante genética (STITZIEL; KIEZUN; SUNYAEV, 2011).

O uso de ferramentas de Bioinformática é essencial para a aplicação de uma técnica como o WES. O desenvolvimento da área da Biologia Molecular trouxe um aumento na complexidade das operações sobre os dados e, assim como, no volume dessas informações. Estratégias de análise de dados incapazes de oferecer algum nível de automação no processo tornam-se inviáveis nesse contexto (PABINGER et al., 2013; SIMS et al., 2014). A maneira como esse instrumental é disponibilizado varia conforme a natureza do problema e depende da infraestrutura computacional acessível ao usuário. Os avanços tecnológicos da Internet nas últimas décadas, com relação à conectividade e à qualidade de experiência do usuário, fizeram dessa plataforma

uma opção para os mais variados serviços, dentre eles os de ferramentas de Bioinformática.

O Exomim é uma aplicação *web* que oferece um conjunto de funcionalidades próprias e integra ferramentas de código aberto de terceiros para a análise de dados em etapas tardias do fluxo de trabalho de um WES. Dentre suas funcionalidades estão a consulta por genes ligados a fenótipos humanos, o armazenamento e gerenciamento de arquivos VCF (*variant call format*) provenientes de plataformas de sequenciamento genético, e a anotação e priorização de variantes também oriundas desse processo. Sua estrutura modular procura permitir que cada uma dessas funcionalidades possa ser utilizada separadamente, conforme a necessidade do pesquisador.

1.1 Motivação

Apesar do enorme avanço na área de pesquisa em Biologia Molecular nas últimas décadas, o acesso a essas novas tecnologias ainda é muito restrito. Um dos principais motivos dessa limitação é o elevado custo financeiro em equipamentos, reagentes, ferramentas de análise, entre outros. Quando consideramos o contexto da pesquisa em universidades federais brasileiras, esse aspecto se torna ainda mais delicado, pois a maior parte do financiamento de projetos é proveniente de verba pública.

Atualmente existem opções de ferramentas com funcionalidades semelhantes às oferecidas pelo Exomim (KOBOLDT et al., 2012; PABINGER et al., 2013). Uma das mais completas é a *variation filter tool*, da Enlis Genomics (ENLIS, 2015a). O custo de uma licença de uso de uma ferramenta varia, podendo chegar a US$ 80,00 por amostra analisada no caso da Enlis (ENLIS, 2015b). Além dos custos, a necessidade de operações personalizadas sobre os dados também estimulam a criação de uma ferramenta capaz de integrar novas funcionalidades para diferentes experimentações sobre os dados de sequenciamento genético.

1.2 Objetivo

Essa aplicação tem como objetivo de suprir uma demanda real dos médicos geneticistas do Hospital de Clínicas de Porto Alegre (HCPA). Sua implementação é uma iniciativa conjunta do Instituto de Informática da Universidade Federal do Rio Grande do Sul (UFRGS) com o HCPA, e busca promover a pesquisa em análise de material genético sequenciado no hospital.

O foco do trabalho é facilitar o diagnóstico clínico em exames de sequenciamento de

regiões importantes do código genético humano, que normalmente estão relacionadas com fenótipos patológicos. Com essa ferramenta, espera-se que a pesquisa em análises clínicas seja impulsionada e que favoreça, principalmente, a população que busca auxílio médico.

1.3 Organização

O restante desse trabalho está organizado da seguinte maneira: No capítulo 2 são apresentados conceitos referentes à genética médica, o estado da arte do sequenciamento de exomas, detalhes sobre a priorização de variantes genéticas e dos pre-requisitos levantados para a criação da aplicação Exomim. No capítulo 3, é descrito o modelo conceitual proposto para o Exomim e seu protótipo implementado, assim como as ferramentas utilizadas no processo de prototipação. No capítulo 4 são apresentadas as etapas de teste e validação do protótipo. Por fim, no capítulo 5 a conclusão do trabalho, com a discussão dos resultados do capítulo 4, propostas para trabalhos futuros e as considerações finais.

2 CONCEITOS E ESTADO DA ARTE

Ao longo desse capítulo serão abordados conceitos da genética médica considerados fundamentais para a compreensão do trabalho. Entre eles destacam-se a importância do aconselhamento genético, os avanços tecnológicos do sequenciamento do DNA e seu papel na identificação de doenças ligadas ao genoma humano, e uma importante metodologia utilizada nesse processo, a priorização de variantes.

A priorização de variantes é uma etapa crucial para que, do material genético sequenciado, tenha-se um diagnóstico clínico de uma mutação patológica. O estado da arte dessa metodologia será tratado com mais detalhes através do levantamento das principais ferramentas de Bioinformática utilizadas atualmente. Nesse contexto será introduzido o Exomim, a aplicação desenvolvida nesse trabalho com o intuito de preencher algumas das lacunas deixadas por essas ferramentas.

2.1 Genética Médica

A genética médica tem como um dos principais objetivos a identificação de alterações genéticas responsáveis pela presença de um fenótipo patológico específico. Sua prática data de meados do século XX, momento histórico onde os avanços científicos e tecnológicos no âmbito molecular começaram a ganhar mais força. Em 1947, a expressão *genetic counseling* (ou aconselhamento genético) foi cunhada pelo médico Sheldon Reed (REED, 1955), que passou a oferecer atendimentos às famílias de pessoas com doenças genéticas. Atualmente, o aconselhamento genético é uma consulta médica que oferece uma importante fonte de informação e, como o nome já diz, aconselhamento a pacientes ou familiares sobre distúrbios genéticos herdados que possam afetá-los ou a seus descendentes. Ele visa o diagnóstico, tratamento e prevenção de doenças genéticas.

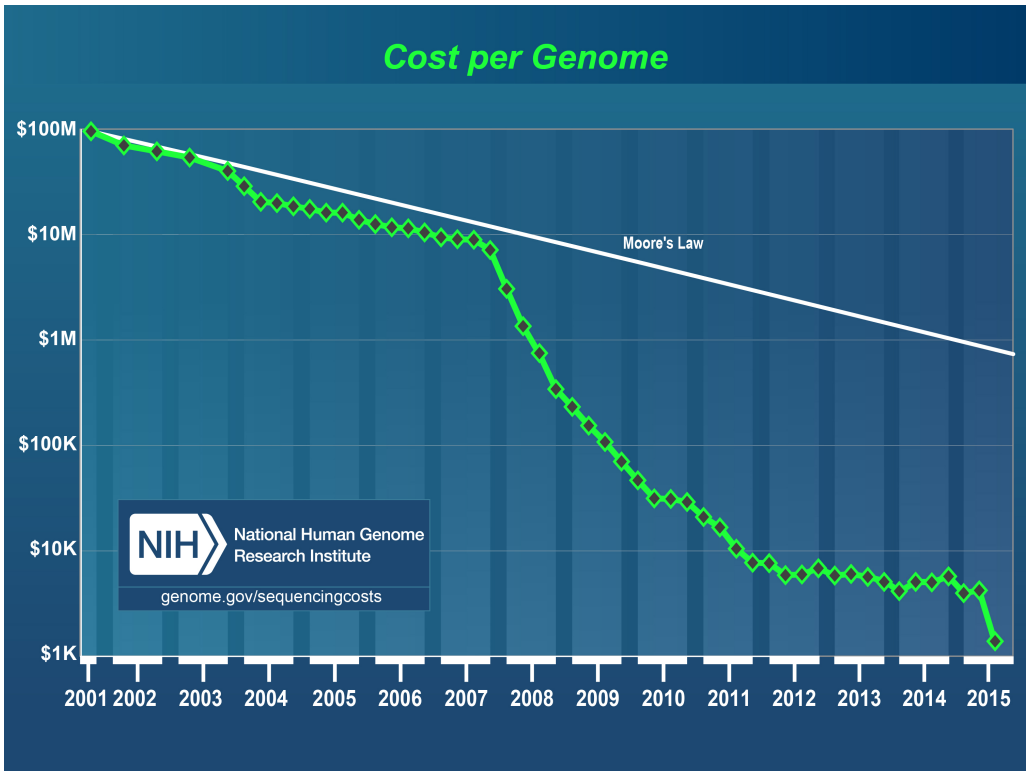
A determinação da mutação causal de uma doença permite a adequação do aconselhamento genético e o correto direcionamento em casos de tratamento. A importância desta questão reflete-se na contínua incorporação de novas tecnologias que possibilitem explicar a base genética de fenótipos específicos de relevância médica (MANOLIO et al., 2009; BAMSHAD et al., 2011; KIEZUN et al., 2012). Essa análise genômica envolve técnicas de citogenética molecular e *next-generation sequencing* (NGS). O NGS representa uma mudança de paradigma na interrogação das variações patogênicas do genoma humano (METZKER, 2010) e vem se tornando a ferramenta de escolha para estudos de genética médica (GOLDSTEIN et al., 2013),

por permitir, entre outras coisas, a análise simultânea de diferentes regiões genômicas.

2.1.1 *Next-Generation Sequencing*

O sequenciamento de ácidos nucleicos é um método para determinar a ordem exata em que nucleotídeos estão organizados em dada molécula de DNA. Na última década, esse sequenciamento tornou-se uma prática acessível e usual a laboratórios de pesquisa ou análises clínicas de todas as partes do mundo. O primeiro grande desafio desse método foi o projeto genoma humano, que foi uma iniciativa internacional e que, ao custo de três bilhões de dólares e duração de 13 anos, completou o sequenciamento de todos os genes do ser humano (genoma) no ano de 2003. A técnica de sequenciamento utilizada nessa ocasião foi o método de Sanger (SANGER; NICKLEN; COULSON, 1977), desenvolvido em 1975 por Edward Sanger e considerado o padrão de sequenciamento por mais de duas décadas.

Figura 2.1 – Capacidade de sequenciamento comparada com a Lei de Moore



Fonte: *National Human Genome Research Institute* (NHGRI)

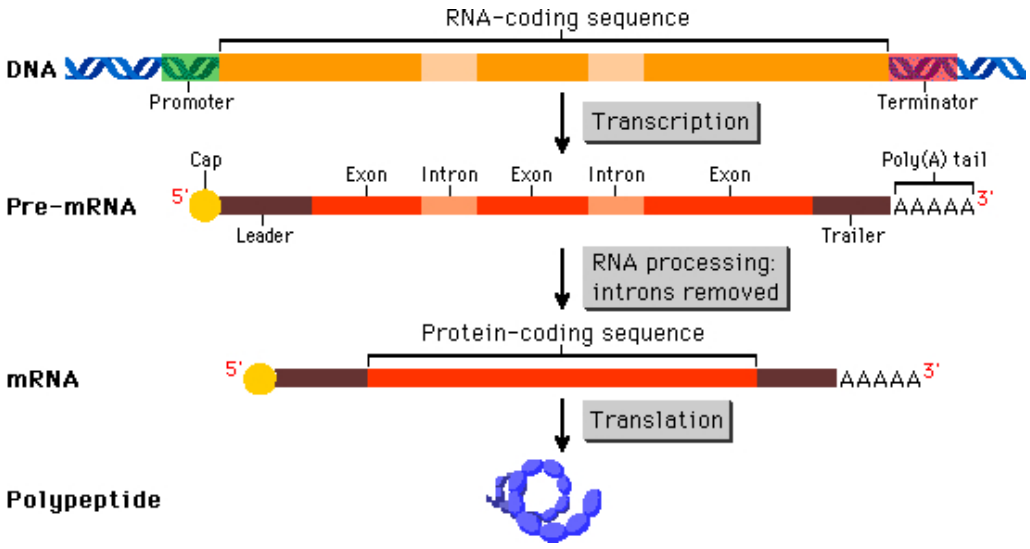
Desde então, a necessidade de métodos de sequenciamento mais rápidos e baratos somente aumentou. Essa demanda guiou o desenvolvimento de novas metodologias, dentre elas o NGS. As plataformas de sequenciamento NGS utilizam uma tecnologia de sequenciamento

massivo paralelo de milhares de fragmentos de DNA, aumentando o *throughput* do sistema e permitindo que um genoma inteiro seja sequenciado em menos de 24 horas (GRADA; WEINBRECHT, 2013). A Figura 2.1 mostra a evolução nos custos envolvidos no sequenciamento do genoma humano. Nela, o custo financeiro (em verde) do sequenciamento é comparado ao ganho em poder computacional estimado pela Lei de Moore (em branco). Acredita-se que a queda abrupta, percebida próxima ao ano de 2007, esteja diretamente relacionada à popularização das plataformas NGS (WETTERSTRAND, 2015).

2.1.2 *Whole-Exome Sequencing*

Dentre as diferentes aplicações para NGS existentes, o *whole-exome sequencing* (WES) é uma das mais utilizadas atualmente. Essa técnica visa o sequenciamento de regiões específicas do genoma, o exoma. O exoma é a parte do genoma formada por éxons, sequências que, quando transcritas, permanecem no RNA maduro após a remoção dos íntrons no processo denominado *splicing*, como apresentado na Figura 2.2, adaptada de *Biology* (MILLER; LEVINE, 2010). A Figura mostra etapas importantes da produção de proteínas realizada no interior das células. A primeira delas, a transcrição do DNA, é o processo onde a informação contida no genoma é copiada para uma nova molécula, considerada uma versão inicial do RNA mensageiro (pre-mRNA). Essa molécula torna-se mRNA após a retirada de partes de sua estrutura (os íntrons), que não formam o conjunto de ácidos ribonucléicos traduzidos ao polipeptídeo final.

Figura 2.2 – Transcrição e tradução do DNA em eucariotos

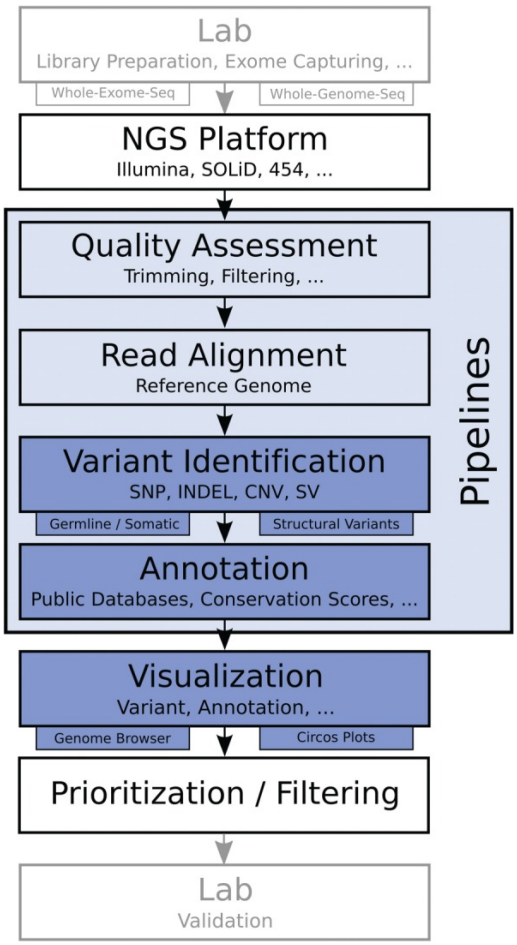


Fonte: Miller & Levine, 2010

O exoma do genoma humano consiste em cerca de 180.000 éxons que constituem de 1-2% do genoma total (NG et al., 2009), mas que acredita-se carregar a maior parte das mutações de efeito patológico (CHOI et al., 2009). Os éxons conhecidos do genoma humano são catalogados e combinados com seus genes equivalentes pelo *Consensus Coding Sequence Database* (CCSD) (COONROD; MARGRAF; VOELKERDING, 2012), um banco de dados de sequências codificantes.

Em um fluxo de trabalho básico para *whole-genome sequencing* (WGS) e WES, as amostras são sequenciadas em uma plataforma NGS e avaliadas conforme sua qualidade e seu alinhamento de leitura contra um genoma de referência. Então, suas variantes são identificadas e as mutações detectadas são anotadas e visualizadas. Por fim, essas mutações podem ser filtradas e priorizadas, e o resultado desse processo validado em laboratório antes de compor um diagnóstico médico, como pode ser visto na Figura 2.3 adaptada de *A survey of tools for variant analysis of next-generation genome sequencing data* (PABINGER et al., 2013).

Figura 2.3 – Fluxo de trabalho básico para WGS e WES



Fonte: Pabinger et al., 2013

Comparado ao WGS, o WES é uma abordagem mais interessante do ponto de vista das análises clínicas. A principal razão para isto é que o conhecimento de variantes genéticas associadas à patogênese de doenças nas regiões codificantes é bastante amplo, enquanto a contribuição das variantes em regiões não-codificantes do genoma é, ainda, limitado (WANG et al., 2013). Ainda, o sequenciamento do exoma provou ser uma estratégia eficiente para determinar a base genética para mais de duas dúzias de distúrbios Mendelianos ou de gene único (BAMSHAD et al., 2011).

Embora as inúmeras aplicações de NGS tenham se tornado uma valiosa fonte de informações biológicas em tempo e custo reduzidos, os desafios no processamento e na interpretação desses dados tornaram a Bioinformática uma área indispensável nesse processo (PABINGER et al., 2013; SIMS et al., 2014). O tratamento do grande volume de dados brutos provenientes dessa análise e a complexidade do refinamento do material processado são alguns dos pontos de maior interesse desse campo científico (NEKRUTENKO; TAYLOR, 2013; MARDIS, 2013). Atualmente, existem uma série de ferramentas disponíveis para as diferentes etapas desse processamento (KOBOLDT et al., 2012; PABINGER et al., 2013); porém, a diversidade de possibilidades de análise desses dados garante o espaço para novas experimentações.

2.2 Priorização de Variantes

Em um fluxo de trabalho de WES, uma vez completo o processo de anotação das variantes, a etapa de priorização pode ser iniciada. A priorização da variante causal entre as milhares encontradas é um passo crucial na análise de exoma, e realizar essa análise, como por exemplo a conexão entre o fenótipo do paciente e o genótipo específico responsável pela condição investigada, é uma tarefa laboriosa (NEKRUTENKO; TAYLOR, 2013). São poucas as ferramentas que englobam estratégias como essa, levando em consideração a relação genótipo/fenótipo para seleção da variante causal entre as mais prováveis (PABINGER et al., 2013). A Tabela 2.1, adaptada de OMIC *Tools* (OMIC, 2015), traz um apanhado de ferramentas que podem ser utilizadas para a priorização de variantes. Cada uma delas apresenta uma abordagem característica mas que, em sua maioria, gira em torno da aplicação de filtros no material anotado do sequenciamento. Das aplicações listadas, destaca-se a eXtasy e o módulo de filtragem da Enlis. O eXtasy é uma dessas ferramentas que procura utilizar o fenótipo como parâmetro de filtragem, entretanto, peca na diversidade dos demais filtros oferecidos. Já o Enlis é uma ferramenta que contempla módulos de análise para diversas etapas do fluxo de trabalho de aplicações NGS, mas é pago, o que dificulta o acesso por parte de diversos grupos de pesquisa a esse instrumento.

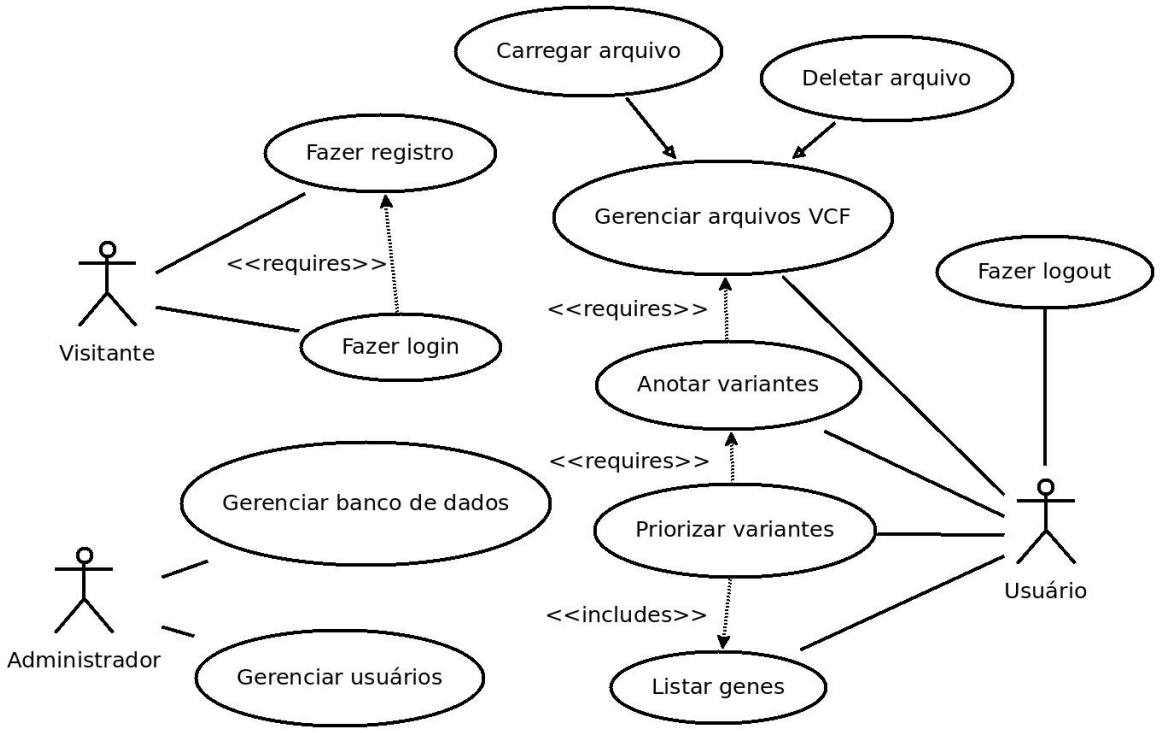
Tabela 2.1 – Ferramentas para priorização de variantes

Nome	*Plataforma*	*Código Aberto*	*Resumo*
BiERapp	Web	Não	Identificação de variantes relacionadas a doenças genéticas esporádicas.
ContrastRank	Web	Não	Priorização especializada para identificação de mutações carcinogênicas.
exomeSuite	Windows	Sim	Filtragem ligada ao genótipo.
ExomeWalker	Web	Não	Identificação de variantes ligadas a doenças Mendelianas.
Exomiser	Web	Não	Anotação funcional de variantes para WES.
eXtasy	Linux, Web	Sim	Variantes de nucleotídeos únicos priorizadas através do fornecimento de fenótipo.
GeneCOST	Linux, Mac, Windows	Não	Priorização baseada em escores de associação a doenças.
KGGSeq	Linux, Mac, Windows	Não	Filtro de três níveis para identificação de variantes Mendelianas.
MendelScan	Linux, Mac, Windows	Não	Basea-se em escore e hereditariedade de variantes.
Olorin	Linux, Mac, Windows	Não	Filtro interativo com seleção de frequência alélica.
OVA	Web	Não	Usa informações fenotípicas usadas para inferir contexto biológico.
Phen-Gen	Linux	Sim	Combina sintomas do paciente para identificar genes em doenças raras.
PhenIX	Web	Não	Priorização a partir de genes candidatos.
PriVar	Linux, Mac, Windows	Não	Identificação de gene candidato e previsão de impacto funcional.
SPRING	Web	Não	Previsão estatística de variante não sinônima responsável pela doença.
Enlis	Mac, Windows, Web	Não	Conjunto de filtros para seleção interativa de variantes.
VAAST	Linux	Não	Busca probabilistica para identificação de genes danificados em genomas.
VAR-MD	Linux	Sim	Preditor de patogenicidade de variantes usando lista ranqueada.
VaRank	Linux	Sim	Anota e ranqueia variantes de acordo com sua patogenicidade.
wKGGSeq	Web	Não	Anota, filtra e prioriza variantes em estudos de WES.

Fonte: OMIC *Tools*

Dentro deste contexto foi elaborada a ferramenta Exomim que foi implementada de maneira modular e focada em uma estratégia especificamente desenhada para abranger os importantes pontos abordados acima. Em linhas gerais, os pré-requisitos do sistema, levantados em conjunto com pesquisadores da área da genética médica do HCPA, são que a ferramenta deve ser capaz de: (1) fornecer informações genotípicas dado um fenótipo patológico; (2) armazenar arquivos de variantes genéticas identificadas no *pipeline* NGS (Figura 2.3); (3) anotar os traços biológicos para cada uma das variantes genéticas presentes nesses arquivos; (4) filtrar o conteúdo das variantes anotadas conforme um conjunto de parâmetros configuráveis por parte do usuário; e (5) suportar a independência entre cada um de seus módulos citados nos itens anteriores. O diferencial do Exomim em relação às demais ferramentas é a incorporação da filtragem por fenótipo junto da oferta de variados filtros para traços biológicos de variantes anotadas. Outra vantagem do sistema é a facilidade na qual ele deve permitir, aos pesquisadores a implementação de novos métodos heurísticos no processo de análise de WES.

Figura 2.4 – Diagrama de casos de uso do Exomim



Para garantir o correto uso de cada uma das funcionalidades do sistema, um conjunto de requisitos secundários precisou ser definido (Figura 2.4), pois seria necessário que se tivesse uma série de cuidados com a privacidade das informações armazenadas nesse sistema para múltiplos usuários. Logo, seria indispensável realizar a autenticação de cada um desses usuários para garantir o mínimo de proteção nesse sentido. Também, o usuário deveria ter uma maneira

de gerenciar os seus documentos, visualizá-los em certos detalhes e apagá-los se fosse preciso. Tendo em vista essas considerações e seguindo uma tendência de ferramentas de Bioinformática (ALEMAN et al., 2014; TIAN; BASU; CAPRIOTTI, 2014; SMEDLEY et al., 2014; ROBINSON et al., 2013; SIFRIM et al., 2013; ANTANAVICIUTE et al., 2015; WU; LI; JIANG, 2014; LI et al., 2015; JAVED; AGRAWAL; NG, 2014), optou-se por uma aplicação *web* com o intuito de oferecer praticidade e disponibilidade ao usuário, sem que esse precisasse se importar com todo o processo de instalação do sistema, e que a única dependência da aplicação fosse uma conexão com a Internet e um navegador.

2.3 Considerações Finais

Nesse capítulo foram apresentados conceitos de genética médica considerados importantes para a compreensão da aplicação proposta nesse trabalho. Foram levantados, também, os principais motivos para a implementação do Exomim, uma ferramenta projetada para auxilio em pesquisas em WES.

O foco do capítulo a seguir está na descrição do projeto e da implementação da aplicação. Será apresentado o resultado da análise de requisitos do sistema através de diagramas de clase e de serviços e será introduzida a arquitetura geral do modelo e serão apresentados os detalhes da prototipação do trabalho.

3 MODELO E PROTÓTIPO

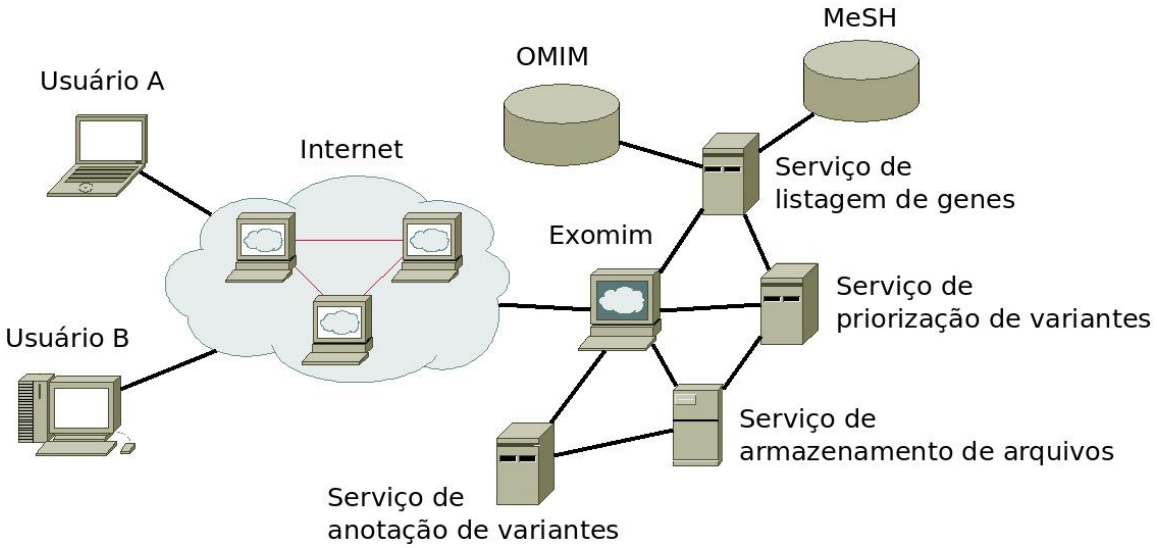
Esse capítulo abordará, em detalhes, o funcionamento dos módulos projetados para o Exomim. Nele serão apresentados o seu modelo teórico, através de uma visão geral dos serviços oferecidos e como eles se relacionam dentro do sistema assim como seu processo de prototipação.

Fez parte da etapa de prototipação a apresentação das ferramentas utilizadas para o desenvolvimento da aplicação. É descrita a maneira como os bancos de dados OMIM (OMIM, 2015a) e MeSH (MESH, 2015a) são integrados no sistema, e como o serviço de listagem de genes utiliza ambos para fazer suas consultas. Em seguida, é detalhado como a ferramenta VEP (HUBBARD et al., 2002) é utilizada na etapa de anotação dos arquivos VCF e, finalmente, o passo a passo do processo de filtragem na priorização de variantes.

3.1 Exomim

O desenvolvimento da aplicação Exomim envolveu etapas de projeto, implementação e análise de uma ferramenta capaz de servir de auxílio em pesquisa e análises clínicas. A criação de seu modelo foi regida pelo conjunto de pré-requisitos estabelecidos pelos pesquisadores do HCPA. Os requisitos tem como objetivo produzir uma aplicação de priorização de variantes adequada para as necessidades dos grupos de pesquisa do hospital.

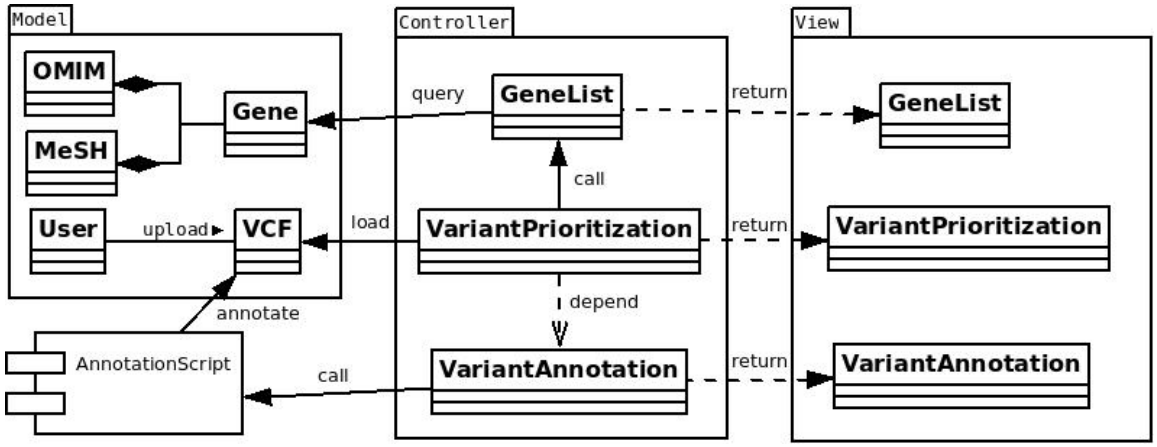
Figura 3.1 – Modelo arquitetural do Exomim



Estruturalmente, o Exomim é dividido em quatro serviços distintos: (1) armazenamento

de arquivos; (2) listagem de genes; (3) anotação de variantes; e (4) priorização de variantes (Figura 3.1). A maneira como cada um desses serviços estão relacionados varia. O serviço de armazenamento de arquivos, por exemplo, é responsável por prover os dados de entrada, tanto para o serviço de anotação, quanto o de priorização. De maneira semelhante, o serviço de listagem de genes pode fornecer uma série de parâmetros extras na priorização de variantes caso o usuário decida acrescentar fenótipos ao processo. A Figura 3.2 apresenta o diagrama de classes do Exomim e mostra em mais detalhes as relações entre essas funcionalidades.

Figura 3.2 – Diagrama de classes do Exomim



O acesso à aplicação ocorre por meio de um navegador e uma conexão com a Internet, no qual o usuário deve se cadastrar para que seus arquivos carregados sejam separados dos demais. No Exomim, os arquivos são particulares, a privacidade deles deve ser garantida pela ferramenta e a interação entre os diferentes serviços deve ocorrer de forma transparente ao usuário. A automação do processo de anotação das variantes genéticas deve ser padronizada para todos os arquivos que buscam a priorização destas na etapa subseqüente. De preferência, deve ser utilizada uma única ferramenta de anotação, pois cada uma carrega um conjunto de particularidades que podem resultar na falta de um dos campos de filtragem do sistema.

A etapa de descrição da prototipação da aplicação retomará questões importantes do modelo. Foi decidido descrevê-lo dessa forma para evitar repetições e para que sua apresentação pudesse ser acompanhada de exemplos.

3.2 Ferramentas Utilizadas

Uma das etapas mais importantes na realização desse projeto foi a escolha do ferramental adequado para a implementação de um protótipo. No que diz respeito a desenvolvimento

web, existem diversas opções para uma série de linguagens de programação, como por exemplo o Laravel e o Yii para desenvolvimento PHP (LARAVEL, 2015; YII, 2015), Ruby on Rails para a linguagem Ruby (RAILS, 2015) e Django, Flask e Pyramid para Python (DJANGO, 2015b; FLASK, 2015; PYRAMID, 2015). Dentre as linguagens de programação citadas, Python é uma das mais utilizadas para o desenvolvimento de aplicações para Bioinformática (MODEL, 2010; KINSER, 2009; STEVENS; BOUCHER, 2015; FLAIG, 2011) e foi a escolhida para a implementação do protótipo da aplicação.

Nas subseções a seguir, são detalhadas as tecnologias adotadas no decorrer do desenvolvimento do projeto. Algumas delas foram de escolha própria do graduando, como no caso do Django e da biblioteca de banco de dados SQLite, mas outras já faziam parte do escopo de pré-requisitos do projeto, como, por exemplo, a ferramenta VEP e o banco de dados OMIM. Para cada uma delas, procurou-se ressaltar os aspectos que foram considerados mais importantes para a compreensão da aplicação Exomim como um sistema.

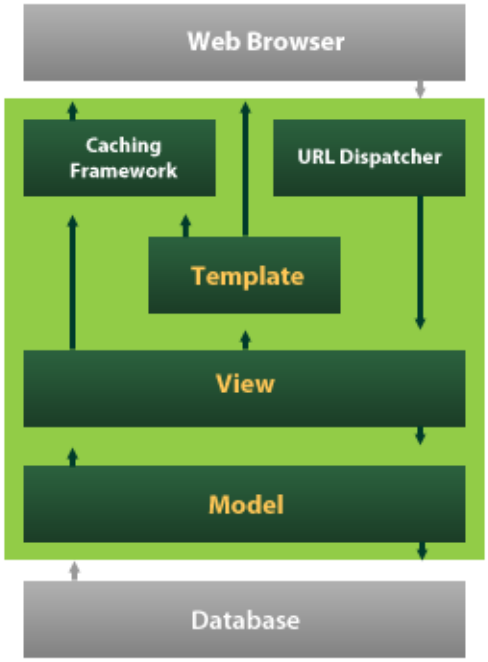
3.2.1 Django

Django é uma ferramenta de propósito geral de aplicações para a Internet, de código aberto e escrito em Python (DJANGO, 2015b). Foi criado no ano de 2003 por um time de desenvolvedores *web* para suprir as necessidades básicas das publicações de um portal de notícias *online*. Sua origem em um ambiente extremamente ativo o moldou para que se tornasse uma ferramenta de fácil manutenabilidade e de boa performance sob carga (KAPLAN-MOSS; HOLOVATY, 2008). Desde 2008 é mantido pela *Django Software Foundation* (DSF), uma organização independente, sem fins lucrativos e que possui uma participativa comunidade de usuários e desenvolvedores (DJANGO, 2015a).

Assim como uma boa quantidade de outras ferramentas, o Django faz uso de um *design pattern* arquitetural similar ao *Model-View-Controller* (MVC), que busca encapsular e isolar os dados e seu processamento (*model*) de sua manipulação (*controller*) e de sua apresentação (*view*). O propósito essencial desse padrão é conectar o modelo mental do usuário com o modelo digital que existe no computador. Seu ideal é promover ao usuário a ilusão de que ele está vendo e manipulando o domínio das informações de maneira direta. Sua estrutura é facilitada se o usuário necessita ver o mesmo modelo simultaneamente em diferentes contextos ou em diferentes pontos de vista (REENSKAUG, 2015). No Django, esse padrão recebe o nome de *Model-Template-View* (MTV), pois um sistema de templates assume o papel da construção das visões e a visão toma conta da manipulação dos dados, tarefa executada pelo controlador no

MVC padrão (DJANGO, 2015c), conforme esquematizado na figura 3.3, adaptada de *Django by example* (WHITTON, 2012).

Figura 3.3 – Diagrama do modelo MTV adotado pelo Django



Fonte: Thomas Whitton, 2012

O *framework* oferece uma série de funcionalidades ao desenvolvedor. Dentre elas, um sistema de *templates* que utiliza o conceito de herança para a criação de páginas HTML, um sistema de internacionalização que facilita a tradução da aplicação para múltiplos idiomas, um sistema extensível de autenticação de usuários e mitigação para diversos ataques típicos de Internet, como a falsificação de solicitação entre sites, *cross-site scripting*, o vazamento de informações de banco de dados através de injeção de SQL, entre outros.

A ferramenta suporta oficialmente quatro *backends* para bancos de dados: PostgreSQL, MySQL, SQLite e Oracle. Entretanto, é possível encontrar extensões para outras bases como Microsoft SQL Server (MANFRE, 2015), IBM DB2 (PRIYADARSHI, 2015), SQL Anywhere (PERROW, 2015) e Firebird (ROBAINA, 2015). Aplicações NoSQL (MongoDB e Google App Engine) também podem ser integradas à ferramenta através da ramificação "django-norel" para bancos de dados não relacionais (DJANGO-NONREL, 2015).

No caso do Exomim, foi utilizada a *engine* de banco de dados SQLite. Como o nome sugere, ela foi desenvolvida para ser uma aplicação leve que se integra ao programa que a utiliza. Diferentemente da maioria dos outros bancos de dados SQL, ela não é montada através de um sistema cliente-servidor. O SQLite lê e escreve os dados diretamente em arquivos no

disco rígido local, onde detém um banco de dados SQL completo com tabelas, índices, *triggers* e *views*. O formato do arquivo da base é multi-plataforma, com compatibilidade entre sistemas de 32 bits e de 64 bits ou até mesmo entre arquiteturas *big-endian* e *little-endian* (SQLITE, 2015).

O SQLite é um banco de dados popular entre diversos tipos de programas, como por exemplo navegadores de Internet, sistemas operacionais e sistemas embarcados. Contém suporte de ligação para várias linguagens de programação. Python, por exemplo, oferece de forma nativa o módulo "sqlite3", que provê a interface de compilação DB-API 2.0 para SQLite (PYTHON, 2015).

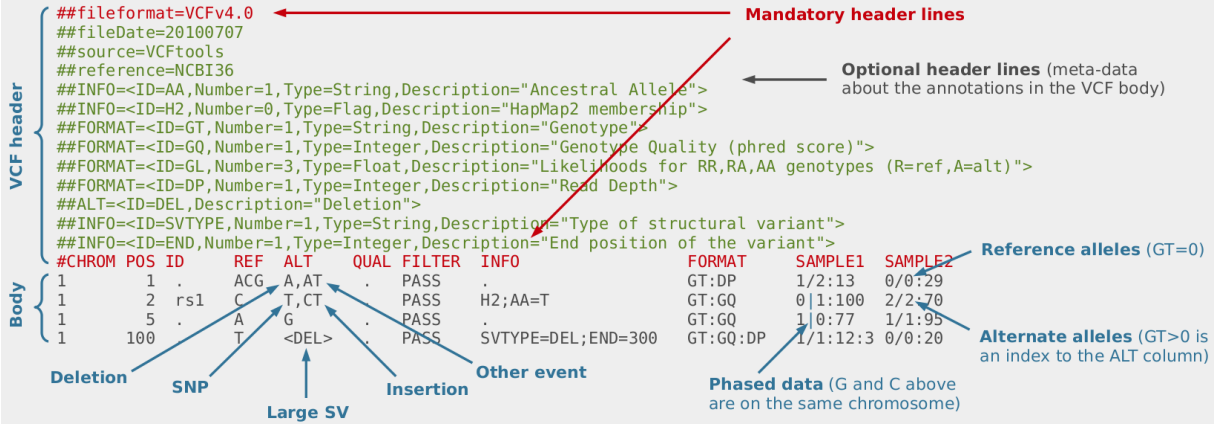
3.2.2 *Variant Effect Preditor*

O *Variant Effect Preditor* (VEP) é uma ferramenta de código aberto mantido pelo Ensembl (ENSEMBL, 2015a), um projeto que disponibiliza de maneira gratuita e online bancos de dados para genomas de vertebrados e outras espécies de seres vivos do domínio *Eukaryota*. A ferramenta integra informações de diversas fontes e promove a anotação de cada uma das variantes presentes em arquivos, por exemplo, do tipo VCF.

O VCF é um formato de arquivo de texto padrão utilizado para armazenar informações de variantes genéticas em diferentes posições cromossômicas (SAMTOOLS, 2015). A Figura 3.4, retirada de *The Variant Call Format and VCFtools* (DANECEK et al., 2015), mostra em detalhes a estrutura de um arquivo VCF. Ele contêm linhas de meta-informações não-obrigatórias, indicadas pela *tag* "##" e uma linha de cabeçalho obrigatória iniciada pela *tag* "#" composta por oito campos fixos, como: (1) o número do cromossomo ao qual pertence a variante (CHROM); (2) a posição da variante em relação ao cromossomo de referência (POS); (3) um identificador único caso a variante já exista em algum banco de dado genético (ID); (4) a base ou o conjunto de bases nitrogenadas de referência para a variante (REF); (5) a base ou o conjunto de bases nitrogenadas alteradas na variante (ALT); (6) o escore confiança da alteração da variante (QUAL); (7) o *status* do processo de filtragem na montagem do arquivo VCF; e (8) um campo para informações adicionais (INFO). As informações a respeito do genótipo da variante também podem ser encontradas na descrição da variante. Esse dado está localizado nos cabeçalhos característicos da amostra, indicado na figura por "SAMPLE1" e "SAMPLE2".

O VCF é montado na etapa denominada "chamada de variante". Essa etapa foge do escopo proposto por esse trabalho e, por isso, pouco foi mencionada ao longo do texto. Ela, assim como a anotação e a priorização de variantes, faz parte do fluxo de trabalho do WES apresentado

Figura 3.4 – O formato VCF



Fonte: Danecek et al., 2015

anteriormente. O produto final desse processo é o arquivo VCF que indica todas as variantes identificadas no sequenciamento genético. Embora esse arquivo já possua informações importantes nesse ponto (i.e. posição, identificação, referência/alteração e genótipo), o significado da variante em um contexto de relevância biológica depende do processo de anotação no qual o VEP se encaixa.

3.2.3 Bancos de Dados

A proposta de listagem de genes do Exomim utiliza dados de dois bancos distintos disponíveis *online*, o MeSH e o OMIM. Os dois bancos são utilizados em conjunto para que se faça uma coleta de genes relacionados a um ou mais fenótipos patológicos humanos. A forma como o Exomim implementa sua listagem de genes é semelhante a modelos já sugeridos (DRIEL et al., 2006) e procura suprir a demanda por uma consulta até então não disponível no OMIM.

O *Medical Subject Headings*, ou simplemente MeSH, é o tesauro controlado pela *National Library of Medicine* (NLM). Ele consiste em um conjunto de descritores para nomeação de termos médicos organizados em uma estrutura hierárquica que permite consultas em diversos níveis de especificidade. Mantido desde 1960, hoje o MeSH conta com um acervo de aproximadamente 28 mil descritores e mais de 87 mil termos auxiliares que facilitam a busca do descritor mais apropriado (MESH, 2015b).

Na Figura 3.5 tem-se um exemplo de um regristro do MeSH. A coluna da esquerda indica o tipo de informação armazenada na coluna da direita. Vale destacar aqui dois desses campos: (1) o *MeSH Heading*, que contém a principal informação de todo o registro que é a

nomenclatura médica oficial para aquele dado; e (2) o campo multi-valorado *Entry Term* que acumula termos considerados sinônimos ao campo descrito no item anterior.

Figura 3.5 – Exemplo de um descritor do MeSH

MeSH Heading	Peritoneum
Tree Number	A01.047.025.600
Tree Number	A10.615.789.596
Annotation	do not confuse with PERITONEAL CAVITY; "peritoneal cells" = probably PERITONEAL CAVITY/cytol or PERITONEAL FLUID (see ASCITIC FLUID)/pathol but not PERITONEUM/cytol; inflammation = PERITONITIS; PERITONEAL LAVAGE is available
Scope Note	A membrane of squamous EPITHELIAL CELLS, the mesothelial cells, covered by apical MICROVILLI that allow rapid absorption of fluid and particles in the PERITONEAL CAVITY. The peritoneum is divided into parietal and visceral components. The parietal peritoneum covers the inside of the ABDOMINAL WALL. The visceral peritoneum covers the intraperitoneal organs. The double-layered peritoneum forms the MESENTERY that suspends these organs from the abdominal wall.
Entry Term	Parietal Peritoneum
Entry Term	Peritoneum, Parietal
Entry Term	Peritoneum, Visceral
Entry Term	Visceral Peritoneum
See Also	Ascitic Fluid
See Also	Laparoscopy
Allowable Qualifiers	AB AH BS CH CY DE EM EN GD IM IN IR ME MI PA PH PP PS RA RE RI SE SU TR UL US VI
Unique ID	D010537

Fonte: MEDLINE *Indexing Online Training Course* (MEDLINE, 2015)

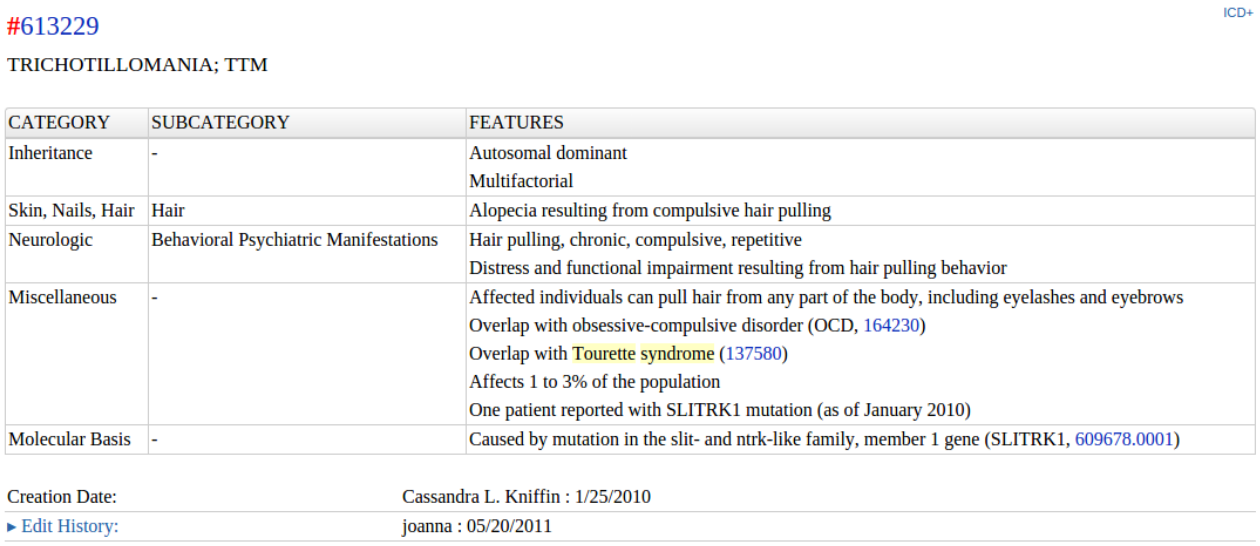
O tesauro do MeSH é utilizado pela NLM na indexação de artigos de mais de cinco mil periódicos biomédicos da base de dados MEDLINE/PubMED, livros, documentos e materiais audiovisuais adquiridos pela entidade. Cada referência bibliográfica da NLM é associada a um conjunto de termos do MeSH que descrevem o conteúdo do item.

A versão impressa do MeSH foi descontinuada no ano de 2007, sendo sua versão online gratuita a única disponível atualmente (MESH, 2015a). Ainda, é possível fazer o *download*, igualmente sem custos, de *snapshots* do banco de dados através da plataforma PubMed (MESH, 2015c) mediante a aceitação dos termos de uso das informações.

Já o *Online Mendelian Inheritance in Man* (OMIM) é um catálogo de genes humanos, traços e distúrbios genéticos, com foco principal na relação entre gene e suas manifestações fenotípicas. Foi criado no ano de 1966 pelo Dr. Victor McKusick, considerado por muitos o "pai da genética médica" (COMFORT, 2012). Naquela época, quando era chamado apenas MIM (*Mendelian Inheritance in Man*), era composto de um compilado de mais de mil descrições de

fenótipos patológicos possivelmente ligados a problemas genéticos (MCKUSICK, 1966). Era publicado com atualizações e correções, em média, de três em três anos até meados da década de 1990. Em 1998, passou por uma grande transformação ganhando espaço na *web* com uma versão *online*, tornando-se a ferramenta como é conhecida hoje, o OMIM.

Figura 3.6 – Registro de uma sinópse clínica do banco de dados OMIM



#613229

ICD+

TRICHOTILLOMANIA; TTM

CATEGORY	SUBCATEGORY	FEATURES
Inheritance	-	Autosomal dominant Multifactorial
Skin, Nails, Hair	Hair	Alopecia resulting from compulsive hair pulling
Neurologic	Behavioral Psychiatric Manifestations	Hair pulling, chronic, compulsive, repetitive Distress and functional impairment resulting from hair pulling behavior
Miscellaneous	-	Affected individuals can pull hair from any part of the body, including eyelashes and eyebrows Overlap with obsessive-compulsive disorder (OCD, 164230) Overlap with Tourette syndrome (137580) Affects 1 to 3% of the population One patient reported with SLITRK1 mutation (as of January 2010)
Molecular Basis	-	Caused by mutation in the slit- and ntrk-like family, member 1 gene (SLITRK1, 609678.0001)

Creation Date:	Cassandra L. Kniffin : 1/25/2010
▸ Edit History:	joanna : 05/20/2011

Fonte: OMIM, 2015

Atualmente armazena mais de 23 mil registros, sendo a maioria deles representando genes relacionados a fenótipos conhecidos (OMIM, 2015a). A descrição detalhada da sinópse clínica representada na Figura 3.6, adaptada de OMIM *clinical synopsis* (OMIM, 2015b), é uma das características principais desse banco de dados. Na imagem vale ressaltar a presença da descrição de sua base molecular com a descrição do gene a qual o fenótipo está relacionado. Esse banco de dados é mantido através de uma parceria da Johns Hopkins University School of Medicine (JHUSOM) com o apoio financeiro do NHGRI e, assim como o MeSH, o OMIM também oferece aos seus usuários, mediante acordo com sua política de uso, o *download* completo de sua base de dados em formato ASCII.

3.3 Listagem de Genes

A listagem de genes é uma funcionalidade que permite ao usuário recuperar os genes relacionados a um ou mais fenótipos humanos. Esse tipo de informação já existe no banco de dados OMIM para traços cuja base molecular é conhecida. Entretanto, a busca por essa

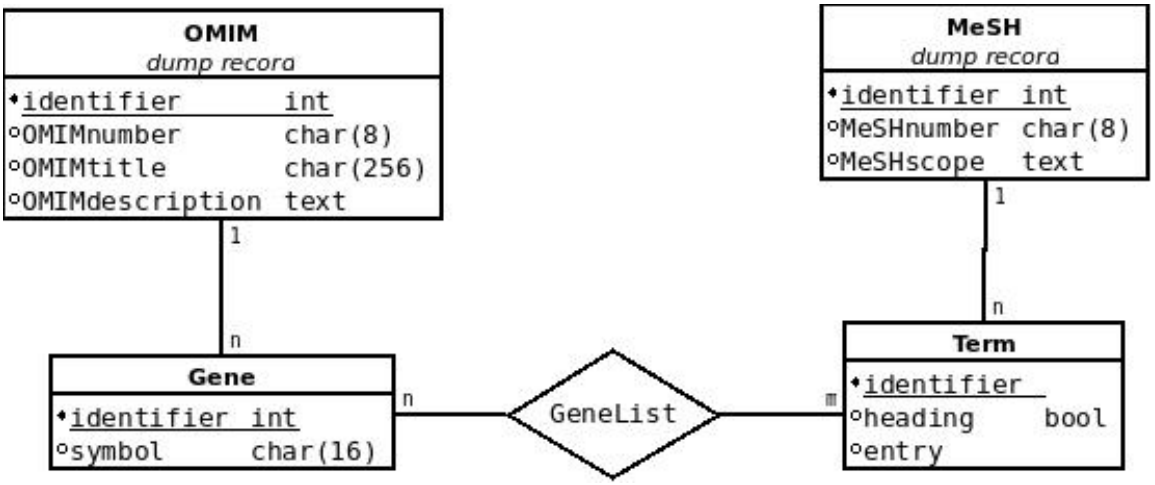
informação requer uma série de consultas consecutivas, a aplicação de restrições específicas nos parâmetros da busca e a filtragem de boa parte dos resultados. A API oferecida pelo banco de dados não provê uma chamada única para que esse tipo de informação seja consultada. Ela, ainda, limita o número de chamadas por segundo que o usuário pode fazer, e o volume de informação contida em cada uma das respostas dessas chamadas. Essa medida é considerada adequada, pois trata-se de um serviço gratuito e que pode ser compartilhado entre centenas de usuários. Porém, torna-se inviável para uma aplicação como a listagem de genes utilizar apenas desse recurso.

A solução para o problema da recuperação dos dados do OMIM foi implantar, localmente, um mecanismo de busca próprio. Para isso, foi necessaŕio fazer o *download* do conteúdo dos registros do OMIM no formato oferecido, uma espécie de *dump* do banco em texto simples e sem associação a qualquer padrão de linguagem estruturada de banco de dados. Para montá-lo localmente, foi preciso a criação de *scripts* com expressões regulares para a coleta das informações de cada um dos registros. Ainda, foi necessário uma forma de pré-processar esse tipo de consulta, pois as informações contidas no OMIM estão dispersas dentro de textos descritivos de variados tamanhos. Para auxílio nesse pré-processamento foi utilizado outro banco de dados, o MeSH. Existe uma preocupação com a atualização dos bancos por isso também foram elaborados *scripts* para a realização dessa tarefa na frequência que o cliente desejar.

Como já abordado no capítulo anterior, o MeSH é um tesauro que contém milhares de termos médicos, incluindo fenótipos humanos. Outro ponto positivo desse banco é que ele armazena, também, um conjunto de sinônimos para cada um desses fenótipos. Dessa forma, é possível pré-processar o conteúdo inteiro do OMIM utilizando como índices os termos médicos oficiais contidos no MeSH e, assim, montar uma tabela com apenas as informações relevantes para uma consulta de listagem de genes.

A Figura 3.7 mostra o diagrama entidade-relacionamento desse mecanismo de busca. As entidades principais, OMIM e MeSH, servem como estruturas para o armazenamento de toda a informação coletada diretamente desses dois bancos. No caso do OMIM, essas informações são o seu identificador original, o título do registro e um compilado de todos os campos descritivos existente no registro original. Cada um dos registros do OMIM é associado a um ou mais genes que possuem um símbolo único. No caso do MeSH, a sua tabela principal contém o identificador original do registro e sua descrição, denominada escopo. As tabelas de termos associadas a cada um dos registros compõem o conjunto de termos médicos e seus sinônimos relacionados àquele escopo. A associação entre a tabela de termos médicos oriunda do MeSH e dos genes do OMIM formam o contexto de consulta da listagem de genes.

Figura 3.7 – Diagrama Entidade Relacionamento



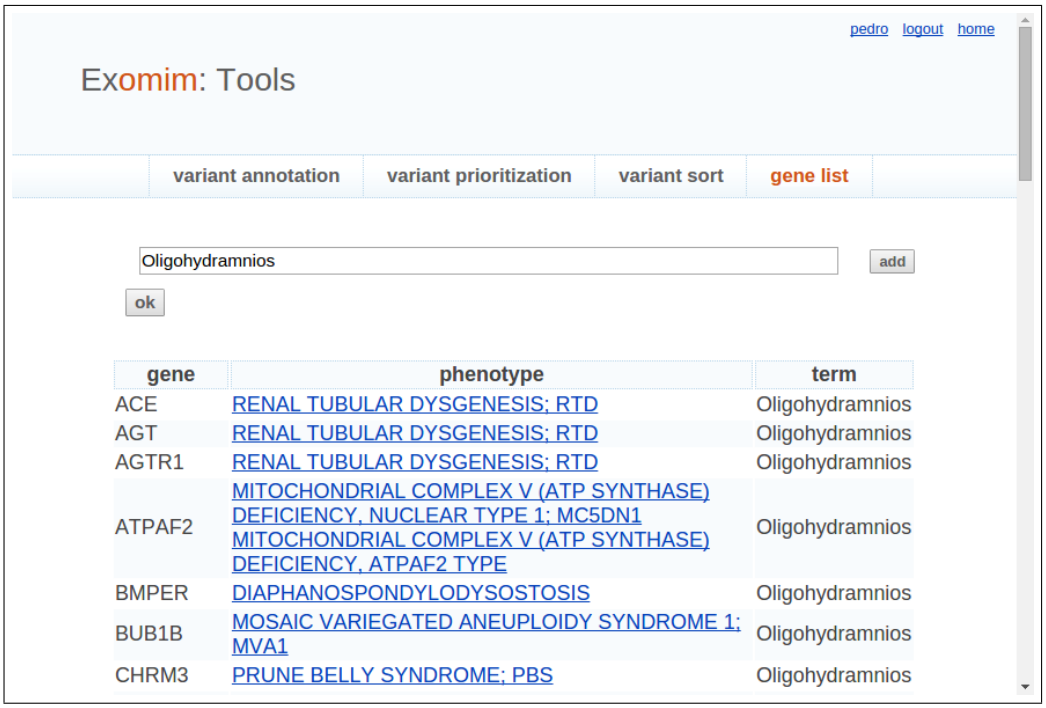
O resultado desse pré-processamento permite que, a partir da entrada de um ou mais fenótipos pelo usuário do Exomim se tenha uma lista de genes associados a esse traço e a todos os seus sinônimos que ocorrem no OMIM. Um exemplo gerado a parir da interface gráfica do Exomim pode ser visto na Figura 3.8. Nela, o usuário solicita todos os genes ligados à "Oligohydramnios", uma situação na gravidez caracterizada pela deficiência de fluido amniótico. O resultado do Exomim é uma lista de entradas do OMIM onde essa condição é mencionada e relacionada a algum gene humano. Cada um dos resultados é ligado a sua referência no OMIM (*hyperlink*) e associado ao termo do MeSH combinado na busca.

3.4 Anotação de Variantes

A anotação das variantes do arquivo VCF, carregado pelo usuário é feita através de uma chamada de execução do *script* VEP (ENSEMBL, 2015a). Além do VEP, outra ferramenta chamada ANNOVAR (ANNOVAR, 2015) foi testada durante a implementação do Exomim. A escolha do VEP se deu devido ao formato da saída do arquivo VFC gerado após a etapa de anotação, que mais assemelhava-se ao formato já utilizado dentro do grupo de pesquisadores do HCPA envolvidos no projeto.

Incluir uma ferramenta de anotação no Exomim foi necessário para oferecer um mecanismo padronizado para o conteúdo das anotações dos arquivos VCF dos usuários. Para que o serviço de priorização funcionasse da maneira esperada, foi preciso que as anotações seguissem um mesmo formato. Assim, através de uma interface gráfica o usuário pode selecionar quais dos seus arquivos (ainda não anotados) e encaminhar um pedido ao servidor para que esse realize a

Figura 3.8 – Tela de consulta de listagem de genes



anotação (Figura 3.9).

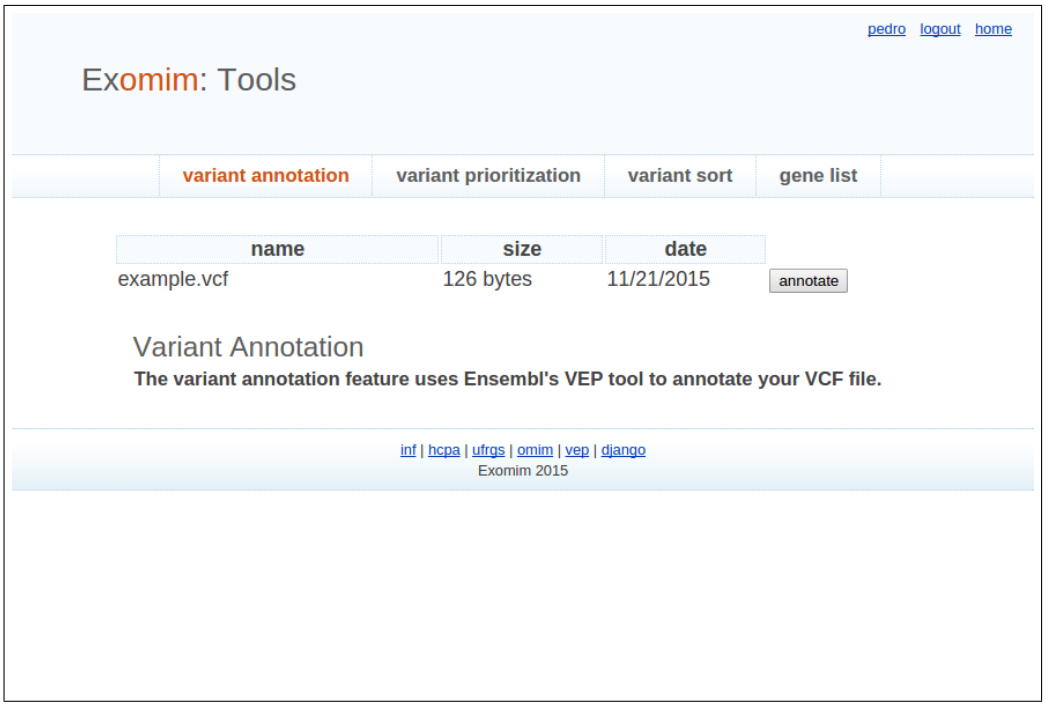
Os parâmetros passados na execução do *script* VEP são fixos, para que todos os arquivos anotados contenham o máximo possível das anotações que passam pelo filtro de priorização. As anotações de interesse de priorização são descritas na Tabela 3.1, adaptada de *Variant Effect Predictor data formats* (ENSEMBL, 2015b).

Tabela 3.1 – Formato dos dados anotados pelo VEP

Cabeçalho da anotação	Descrição do formato
Consequence	Tipo de consequência da variante
IMPACT	O modificador de impacto para o tipo de consequência
SYMBOL	O símbolo do gene
Feature type	Tipo de característica: *Transcript, RegulatoryFeature or MotifFeature*
BIOTYPE	Biotipo do transcrito ou da característica regulatória
Existing_variation	Identificadores de variações COSMIC e dbSNP
VARIANT_CLASS	Classe de ontologia da variante
SIFT	Predição e/ou escore SIFT
PolyPhen	Predição e/ou escore PolyPhen
GMAF	Frequência global da variante existente no *1000 Genomes*
CLIN_SIG	Significância clínica da variante do dbSNP

Fonte: VEP

Figura 3.9 – Tela de anotação de variantes



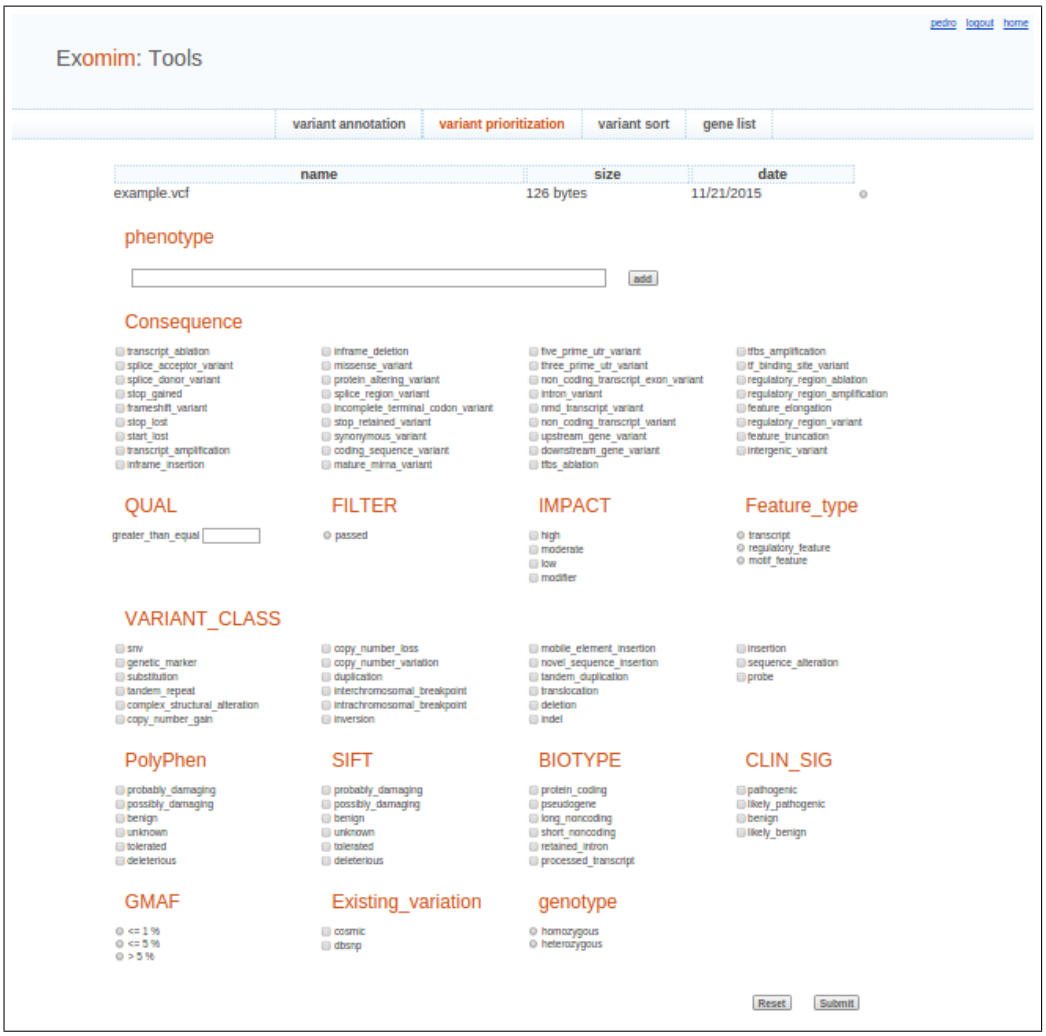
3.5 Priorização de Variantes

No Exomim, as variantes de um arquivo VCF anotado são priorizadas conforme um conjunto de filtros. Essa informação é coletada através de uma interface gráfica que oferece todas as opções de filtros ao usuário. A Figura 3.10 apresenta a tela de priorização da aplicação em tamanho reduzido, e serve para que se tenha uma dimensão da quantidade de alternativas na priorização de um arquivo VCF. Os filtros utilizados nesse módulo foram selecionados por pesquisadores da genética médica do HCPA.

Toda a informação de entrada do usuário é repassada ao módulo de priorização, que irá validar cada um dos filtros e aplicá-los a cada uma das variantes. A etapa de validação ocorre a partir da leitura dos metadados do arquivo VCF. Nele estão contidos os cabeçalhos que indicam todos os dados existentes para cada uma das variantes. Os cabeçalhos de interesse do Exomim são o das anotações do VEP e os cabeçalhos obrigatórios do arquivo VCF (*mandatory header lines*), já vistos na Figura 3.4. Os metadados oriundos do processo de anotação podem variar conforme os parâmetros passados ao *script* do VEP antes de sua execução. Como visto na seção anterior, o Exomim utiliza o VEP com parâmetros fixados para que todas as variantes anotadas pela aplicação mantenham o mesmo formato. Embora esse cuidado ofereça uma maior consistência aos mecanismos de priorização, ele não garante que o VEP terá toda a informação

necessária para anotar todos os campos de cada uma das variantes, até porque nem todas as mutações do exoma humano são conhecidas. Nesse caso, o sistema de filtragem fica responsável por garantir a integridade do processo de priorização, evitando o descarte de variantes que, por algum motivo, não possuam dados referentes a algum dos filtros do usuário.

Figura 3.10 – Tela de priorização de variantes



A ordem preferencial de filtragem é da aplicação dos elementos considerados mais excludentes aos menos excludentes. O objetivo dessa organização é reduzir o custo computacional do processo, e nessa implementação os filtros considerados mais restritivos são aqueles que oferecem o menor número de opções de filtragem e, de preferência, que a escolha de uma opção exclua qualquer variante que contenha a outra, por exemplo o *genotype*. A adoção do *genotype* na priorização implica que o usuário vá filtrar apenas variantes pertencentes a alelos homozigotos ou pertencentes a alelos heterozigotos.

Um dos diferenciais do Exomim é a integração do serviço de listagem de genes com

o módulo de priorização. Dessa forma é possível inserir, além dos filtros padrões, traços fenotípicos que possam caracterizar o paciente e, talvez, auxiliar no isolamento de uma variante genética específica.

3.6 Considerações Finais

Nesse capítulo foram descritas as etapas de projeto e implementação do Exomim. O objetivo na elaboração do modelo e na construção do protótipo foi traduzir em uma ferramenta os requisitos de um sistema de priorização de varientes mais adequado às necessidades de pesquisadores da genética médica do HCPA.

O capítulo seguinte atenderá a etapa de análise das funções consideradas fundamentais do Exomim: a listagem de genes, o uso dos filtros na priorização de variantes e a experiência de usuário no uso de todas as funcionalidades oferecidas.

4 EXPERIMENTOS E RESULTADOS

Nesse capítulo estão descritos os processos de teste e os resultados obtidos com análises feitas para algumas das funcionalidades do Exomim. Essas funcionalidades foram escolhidas pois acredita-se serem as de maior relevância para esse trabalho. Todos os experimentos foram realizados utilizando um servidor local, implantado em um ambiente virtualizado. Nem todos os resultados estão descritos nesse capítulo. A seção de apêncides apresenta todo o material extra produzido na realização desses testes.

Os experimentos estão divididos de acordo com sua especificidade. O primeiro deles, a listagem de genes, procurou investigar o quão semelhante os resultados dessa consulta estão com os resultados obtidos a partir de consultas no OMIM. O segundo, a priorização de variantes, explorou o uso dos filtros em arquivos VCF reais anotados no próprio Exomim. Por fim, o terceiro experimento buscou fazer uma síntese de um estudo de caso em experiência de usuário no uso da ferramenta.

4.1 Listagem de Genes

Por se tratar de um mecanismo de busca que procura o máximo de similaridade possível ao mecanismo de busca original foi necessário realizar um experimento para avaliar o quão próxima a listagem de genes realizada no Exomim está dos dados obtidos através de uma busca no OMIM utilizando um mesmo parâmetro. Os parâmetros utilizados nesse experimento foram obtidos juntamente com pesquisadores do HCPA. Trata-se de fenótipos clínicos associados a doenças de herança mendeliana, selecionados a partir de consulta à literatura médica.

Para cada um dos fenótipos foi realizada uma consulta para *clinical synopsis* na plataforma *online* do OMIM. O filtro utilizado foi o de que somente deveriam ser selecionados registros cuja base molecular estivesse descrita. Esse tipo de filtragem é importante pois é no campo da base molecular onde estão contidos os dados a respeito dos genes relacionados ao registro. Como resposta, o OMIM gera uma lista de identificadores (*MIM Number*) de seus registros que estão associados aos parâmetros da busca. No lado da consulta da listagem de genes, foi realizada uma busca para cada um dos fenótipos selecionados e feita a coleta da referência do *MIM Number* de cada um dos genes listados para aquele dado fenótipo.

Os resultados obtidos foram compilados na Tabela 4.1. Nela, a primeira coluna informa o fenótipo de entrada utilizado seguido do número de registros obtidos pelo mecanismo de busca do OMIM (A) e o número de registros provenientes do mecanismo de busca da listagem

de genes do Exomim(B). A quarta coluna apresenta a quantidade de registros presentes na intersecção dessas duas consultas ($|A \cap B|$) e, por fim, a útlima coluna é o coeficiente de similaridade de Jaccard calculado para o experimento. Esse coeficiente foi escolhido pois ele leva em consideração, além da similaridade, a diversidade dos conjuntos para estabelecer seu valor, um número que varia de 0 à 1, onde 1 é equivalência total dos conjuntos. A média calculada para o coeficiente de similaridade de Jaccard para as 15 consultas foi de $\overline{J(A,B)} = 0.798$. O coeficiente de similaridade de Jaccard varia entre valores próximos de zero até um. Quanto maior o valor maior a similaridade entre os dois conjuntos analisados.

Tabela 4.1 – Resultado do experimento de validação da listagem de genes

Fenótipo	$\|A\|$	$\|B\|$	$\|A \cap B\|$	$J(A,B)$
dyskinesias	82	33	33	0.402
epistaxis	45	45	45	1.000
anemia	200	272	192	0.686
angiokeratoma	11	9	9	0.818
cardiomyopathy	200	260	194	0.729
cholestasis	62	59	59	0.952
demyelination	92	89	89	0.967
malnutrition	27	26	26	0.963
pancytopenia	64	60	60	0.938
hemochromatosis	22	21	21	0.955
hepatomegaly	200	243	196	0.794
hyperreflexia	200	282	196	0.685
jaundice	78	77	76	0.962
leukodystrophy	62	24	15	0.211
stroke	105	99	97	0.907

4.2 Priorização de Variantes

Os experimentos realizados para validar o módulo de priorização de variantes foram planejados com o intuito de testar o correto funcionamento dos filtros propostos. O material utilizado para esse procedimento foi cedido por pesquisadores do HCPA. Ao todo, foram utilizadas 35 amostras de arquivos VCF contendo, cada uma, um conjunto de variantes de material genético humano sequenciado.

Primeiramente, todos os arquivos foram carregados para o Exomim através de seu serviço de gerenciamento de arquivos e, um a um, foram anotados utilizando o serviço de anotação de variantes. O conjunto de filtros utilizados para cada uma das amostras foi definido de forma que o resultado final do crivo fosse de apenas uma variante. Procurou-se, desse modo, tes-

tar o maior número possível de filtros em todos os casos. Entretanto, nem todas as variantes possuíam dados biológicos anotados para cada um dos campos filtrados.

A Tabela 4.2 resume, por motivo de espaço, o conjunto de filtros utilizados para 25 das 35 amostras. O nome de cada um dos filtros foi reduzido para suas quatro primeiras letras. São omitidos os filtros externos ao conteúdo da anotação: qualidade da amostra (QUAL); status da filtragem da etapa de identificação das variantes (FILTER); e o genótipo da variante (genotype). Esses três filtros foram utilizados para todas as amostras, pois são independentes do resultado da etapa de anotação. O filtro SYMBOL não foi utilizado para esse experimento, e a razão disso é que ele é utilizado indiretamente pelo módulo de priorização. A validação desse filtro está relacionada com a validação do serviço de listagem de genes, pois os genes que compõem os elementos dessa filtragem são provenientes de consultas por fenótipos. Os resultados obtidos são os esperados para essa funcionalidade.

Em média, cada amostra continha cerca de 15 variantes. Para todos os casos testados foi possível reduzir o número de variantes da amostra a apenas uma. O resultado do restante das amostras desse experimento está localizado na seção de apêndices desse trabalho.

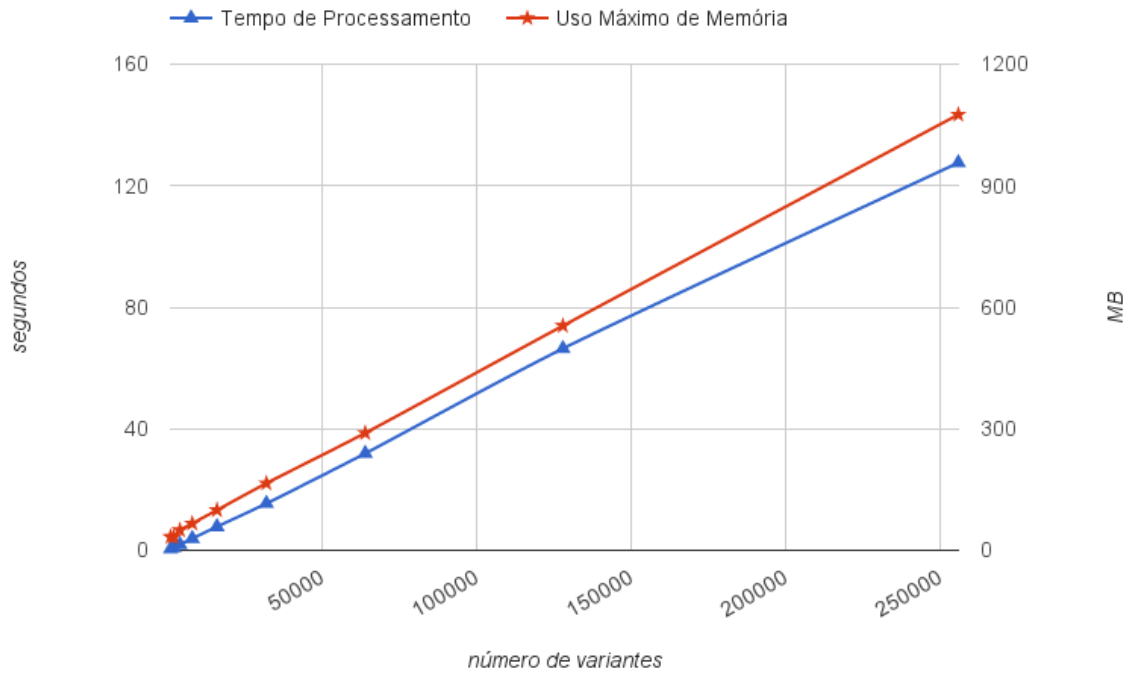
Para se ter uma estimativa do desempenho do módulo de priorização de variantes, foram realizados experimentos onde variou-se o tamanho da entrada. As entradas foram geradas a partir de um compilado de arquivos VCF anotados, disponibilizados por pesquisadores do HCPA. Ao todo, esses arquivos compuseram um só VCF com aproximadamente 1000 variantes e esse arquivo foi replicado para a geração das outras entradas. As quantidades de variantes de entradas utilizadas foram de 1000, 2000, 4000, 8000, 16000, 32000, 64000, 128000 e 256000. O tamanho do arquivo com 256000 variantes tinha, aproximadamente, 1 *gigabyte*. O experimento foi conduzido em um computador com processador Intel® Core™ i7-4500U com 7,7 GiB de memória RAM. O servidor da aplicação estava em um container Linux com sistema operacional Ubuntu 14.04-LTS.

Nos experimentos foram mensurados o tempo de processamento e o máximo de memória gasto pela aplicação. Para a captura do tempo de processamento, foi utilizada a biblioteca *time* de Python e para registrar o consumo de memória do processo foi utilizado o comando *top* do terminal do Linux. A Figura 4.1 apresenta os valores médios calculados para 5 repetições de cada um dos experimentos. Na seção de apêndices estão localizadas as tabelas contendo o valor obtido para cada um dos experimentos, assim como a média aritimética calculada para cada um deles.

Tabela 4.2 – Resultado do experimento de validação da priorização de variantes

Amostra	*Número de variantes totais*	*Filtros utilizados*									
		Cons	*IMPA*	*Feat*	*BIOT*	*Exis*	*VARI*	*SIFT*	*Poly*	*GMAF*	*CLIN*
1	14	X	X	X	X	X	X	-	-	-	-
2	20	X	X	X	X	-	X	-	-	-	-
3	14	X	X	X	X	X	X	-	-	X	X
4	13	X	X	X	X	X	-	-	-	X	X
5	12	X	X	X	X	X	X	X	X	X	X
6	18	X	X	X	-	X	X	-	-	X	-
7	13	X	X	X	X	X	X	-	-	X	-
8	12	X	X	X	X	X	X	-	-	X	-
9	12	X	X	X	X	-	X	X	X	X	X
10	12	X	X	X	X	X	X	-	-	X	X
11	21	X	X	X	X	X	X	-	-	X	X
12	8	X	X	X	X	-	X	-	-	-	-
13	24	X	X	X	X	X	X	-	-	X	X
14	4	X	X	X	X	X	X	-	-	X	-
15	25	X	X	X	X	X	X	X	X	X	X
16	11	X	X	X	X	-	X	X	X	-	-
17	22	X	X	X	X	-	X	-	-	-	-
18	10	X	X	X	-	X	X	-	-	X	X
19	20	X	X	X	-	X	X	-	-	X	X
20	23	X	X	X	X	X	-	-	-	X	-
21	9	X	X	X	-	X	X	-	-	X	X
22	9	X	X	X	X	X	X	-	-	X	-
23	10	X	X	X	-	X	X	-	-	X	X
24	11	X	X	X	X	X	X	-	-	-	-
25	9	X	X	X	X	X	X	-	-	X	-

Figura 4.1 – Perfil de desempenho do módulo de priorização de variantes



4.3 Experiência do Usuário

Para completar a etapa de resultados da aplicação, foi conduzido um experimento piloto de caráter qualitativo de interação com usuário. O objetivo dessa avaliação foi registrar a experiência do usuário em um primeiro contato com as interfaces e funcionalidades que o sistema oferece no que diz respeito à correção dos resultados. O experimento foi dividido em 5 etapas: (1) registro de novo usuário; (2) *upload* e visualização de arquivos na pasta do usuário; (3) anotação de variantes; (4) listagem de genes; e (5) priorização de variantes.

O participante selecionado para o experimento faz parte do grupo de pesquisadores da genética médica do HCPA e da pós-graduação em genética da UFRGS. Um dos aspectos explorados nessa atividade foi a existência da experiência prévia desse usuário com ferramentas de Bioinformática dessa natureza.

A condução do experimento deu-se no estabelecimento de tarefas ao usuário que contemplassem a realização das 5 etapas de avaliação descritas anteriormente. Após a finalização de cada uma das etapas, foram registradas as impressões positivas e negativas, assim como foi

pedido ao usuário que fizesse uma breve avaliação comparativa com algum outro sistema de mesmo propósito.

As etapas iniciais do experimento foram focadas nas tarefas básicas da ferramenta. Após o registro de uma nova conta de usuário, o participante foi convidado a fazer o *upload* de um arquivo VCF para a continuação da atividade. Essas etapas geraram poucos comentários por parte do usuário. Porém, a execução do procedimento de anotação das variantes, presentes no arquivo VCF carregado na etapa anterior gerou um descontentamento do usuário devido à demora do processo. O participante estranhou o tempo levado para que a ferramenta VEP completasse a anotação de seu arquivo.

Na avaliação do módulo de listagem de genes, o usuário teve a liberdade de buscar qualquer termo médico, com a única restrição de que esse deveria ser no idioma inglês. Segundo o participante, essa foi a tarefa que mais lhe surpreendeu. Nas suas palavras, a simplicidade e facilidade na qual a consulta é realizada difere da maneira como outras ferramentas procuram realizar a mesma tarefa.

Por fim, na etapa de priorização de variantes, o usuário relatou que teve dificuldade na seleção dos filtros. Ele apontou que, dos filtros oferecidos para a priorização, muitas das opções possuem um caráter associativo. Por exemplo, o fator de impacto da variante (IMPACT) está diretamente relacionado com o fator consequência. Logo, a escolha ou exclusão de fatores de impacto deveriam ser responsivos às consequências dispostas na interface. Esse problema pode ser corrigido melhorando a disposição dos elementos da tela e adotando interfaces responsivas.

4.4 Considerações Finais

Os experimentos e resultados apresentados nesse capítulo procuraram avaliar o protótipo e suas principais funcionalidades da maneira como eles existem hoje. A ideia principal na elaboração de cada um foi que, desse momento em diante pudesse ser direcionado os próximos passos desse projeto.

O capítulo seguinte entrará com uma discussão um pouco mais aprofundada dos resultados obtidos e fará uma síntese do trabalho realizado até aqui. Ainda, serão propostos trabalhos futuros a serem realizados com essa aplicação.

5 CONCLUSÃO

Ao longo desse trabalho foram descritas as etapas de projeto, implementação e análise de uma aplicação *web* para priorização de variantes em *whole-exome sequencing*. 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.

O valor obtido como média do coeficiente de similaridade de Jaccard, na comparação dos resultados da busca do OMIM com os resultados da listagem de genes, mostra que é possível explorar uma estratégia como essa para preencher as lacunas deixadas pela API do OMIM. Quando analisamos os valores de Jaccard obtidos individualmente, temos um contexto de similaridade maior que 0.9 para mais da metade dos fenótipos testados. Entretanto, a presença de valores considerados baixos, observados em alguns casos, indicam diferentes situações que devem ser analisadas futuramente:

1. Desses registros discrepantes, quais realmente são condizentes conteúdo da busca? Nos experimentos assumimos que o OMIM é capaz de realizar a melhor consulta possível sobre seus dados, porém, na listagem de genes também são utilizados termos sinônimos que podem não ser indexados pelo motor de busca original;
2. A listagem de genes retorna apenas registros que, de fato, apresentam pelo menos um gene associado à base molecular do registro do OMIM. Contudo, o resultado da busca do OMIM não tem essa finalidade e a presença de registros cuja base molecular não contém genes, mas cromossomos inteiros por exemplo, pode ser o motivo da diferença em alguns casos;
3. Por trás do motor de busca do OMIM pode existir um mecanismo não previsto no projeto da listagem de genes do Exomim que está fazendo a diferença na coleta da informação.

Os filtros do módulo de priorização de variantes apresentaram o comportamento esperado nos experimentos com os arquivos VCF fornecidos pelos pesquisadores do HCPA. Em todos os casos, o uso dos filtros foi capaz de reduzir o conjunto de variantes de cada um desses arquivos a uma só. Espera-se que sejam incluídos novos campos de filtragem a medida que essa aplicação for se desenvolvendo. Essas inclusões devem estar companhadas de novos testes semelhantes a esse para que se possa garantir o funcionamento desse serviço.

Ainda sobre o módulo de priorização de variantes, seu desempenho apresentou um ganho linear no consumo de memória e tempo de processamento conforme a variação no tamanho

da entrada. Esse fator é importante, por exemplo, em situações onde mais de um usuário estará compartilhando a aplicação simultaneamente. Recomenda-se, para trabalhos futuros, que sejam realizados testes de carga em ambientes que simulem a presença de múltiplos usuários.

O experimento piloto de experiência com o usuário foi importante para a identificação dos pontos onde a aplicação deve melhorar. O módulo de anotação de variantes precisa ser revisto para que seja confirmado se, de fato, existe alguma demora por parte da ferramenta de anotação. A listagem de genes mostrou-se de grande importância ao sistema e, unida ao serviço de priorização, pode ser o grande diferencial do Exomim.

Espera-se que haja a continuidade desse trabalho. Existe um longo caminho pela frente para que o Exomim deixe de ser um protótipo e passe a ser um produto final. O aprimoramento do mecanismo de listagem de genes deve ser um dos focos para trabalhos futuros em cima dessa aplicação, assim como o aperfeiçoamento de sua interface gráfica. Além disso, novas funcionalidades devem ser integradas à ferramenta para que ela torne-se ainda mais completa.

Os resultados obtidos através dos experimentos realizados com o protótipo do Exomim justificam os esforços na realização desse trabalho. A aproximação à multidisciplinaridade dentro do ambiente acadêmico é algo pouco explorado em diversas áreas do conhecimento. A Bioinformática abre um canal de comunicação entre a ciência da computação e as ciências biológicas e produz resultados de extrema relevância ao meio científico. Espero que esse trabalho também sirva como incentivo para projetos que envolvendo diferentes grupos de pesquisa continuem acontecendo e que ganhem ainda mais espaço dentro da universidade.

REFERÊNCIAS

ALEMAN, A. et al. A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. **Nucleic Acids Research**, 2014.

ANNOVAR. **ANNOVAR Documentation**. 2015. Disponível em: <http://annovar.openbioinformatics.org/en/latest/>. Acessado: 22-11-2015.

ANTANAVICIUTE, A. et al. Ova: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. **Bioinformatics**, 2015.

BAMSHAD, M. J. et al. Exome sequencing as a tool for Mendelian disease gene discovery. **Nature Reviews Genetics**, 2011.

CHOI, M. et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. **Proceedings of the National Academy of Sciences**, 2009.

COMFORT, N. **The Science of Human Perfection**. [S.l.]: Yale University Press, 2012.

COONROD, E. M.; MARGRAF, R. L.; VOELKERDING, K. V. Translating exome sequencing from research to clinical diagnostics. **Clinical Chemistry and Laboratory Medicine**, 2012.

DANECEK, P. et al. **The Variant Call Format and VCFtools**. 2015. Disponível em: <http://vcftools.sourceforge.net/VCF-poster.pdf>. Acessado: 15-11-2015.

DJANGO. **Django Code**. 2015. Disponível em: <https://github.com/django/django>. Acessado: 14-11-2015.

DJANGO. **Django Project**. 2015. Disponível em: <https://www.djangoproject.com>. Acessado: 14-11-2015.

DJANGO. **FAQ: General**. 2015. Disponível em: <https://docs.djangoproject.com/en/1.8/faq/general/>. Acessado: 22-11-2015.

DJANGO-NONREL. **Making Django run on non-relational databases**. 2015. Disponível em: <http://django-nonrel.org/>. Acessado: 14-11-2015.

DRIEL, M. A. van et al. A text-mining analysis of the human phenome. **European Journal of Human Genetics**, 2006.

ENLIS. **Enlis Genomics**. 2015. Disponível em: <https://www.enlis.com/>. Acessado: 22-11-2015.

ENLIS. **Frequently asked questions**. 2015. Disponível em: <https://www.enlis.com/import/#collapse4>. Acessado: 19-12-2015.

ENSEMBL. **Ensembl tools including the VEP**. 2015. Disponível em: <https://github.com/Ensembl/ensembl-tools>. Acessado: 15-11-2015.

ENSEMBL. **Variant Effect Predictor data formats**. 2015. Disponível em: <http://www.ensembl.org/info/docs/tools/vep/vep_formats.html>. Acessado: 22-11-2015.

FLAIG, R. M. **Bioinformatics Programming in Python: A Practical Course for Beginners**. [S.l.]: Wiley-Blackwell, 2011.

FLASK. **Web development, one drop at a time**. 2015. Disponível em: <http://flask.pocoo.org/>. Acessado: 18-11-2015.

GOLDSTEIN, D. B. et al. Sequencing studies in human genetics: design and interpretation. **Nature Reviews Genetics**, 2013.

GRADA, A.; WEINBRECHT, K. Next-Generation Sequencing: Methodology and Application. **Journal of Investigative Dermatology**, 2013.

GRIFFITHS, A. J. F. et al. **Introduction to Genetic Analysis**. [S.l.]: W.H. Freeman, 2012.

HUBBARD, T. et al. The ensembl genome database project. **Nucleic Acids Research**, 2002.

JAVED, A.; AGRAWAL, S.; NG, P. C. Phen-gen: combining phenotype and genotype to analyze rare disorders. **Nature Methods**, 2014.

KAPLAN-MOSS, J.; HOLOVATY, A. **The Definitive Guide to Django: Web Development Done Right**. [S.l.]: Apress, 2008.

KIEZUN, A. et al. Exome sequencing and the genetic basis of complex traits. **Nature Genetics**, 2012.

KINSER, J. **Python for Bioinformatics**. [S.l.]: Jones and Bartlett, 2009.

KOBOLDT, D. C. et al. Massively parallel sequencing approaches for characterization of structural variation. **Methods in Molecular Biology**, 2012.

LARAVEL. **The PHP Framework For Web Artisans**. 2015. Disponível em: <http://laravel.com/>. Acessado: 18-11-2015.

LI, M. J. et al. wkggseq: A comprehensive strategy-based and disease-targeted online framework to facilitate exome sequencing studies of inherited disorders. **Human mutation**, 2015.

MANFRE, M. **Django MSSQL Database Backend**. 2015. Disponível em: <https://bitbucket.org/Manfre/django-mssql/src>. Acessado: 14-11-2015.

MANOLIO, T. A. et al. Finding the missing heritability of complex diseases. **Nature**, 2009.

MARDIS, E. R. Next-generation sequencing platforms. **Annual Review of Analytical Chemistry**, 2013.

MCKUSICK, V. A. **Mendelian Inheritance in Man. Catalogs of Autosomal Dominant, Autosomal Recessive and X-Linked Phenotypes**. [S.l.]: Johns Hopkins University Press, 1966.

MEDLINE. **MEDLINE Indexing Online Training Course**. 2015. Disponível em: <https://www.nlm.nih.gov/bsd/indexing/training/MSH_030.html>. Acessado: 13-11-2015.

MESH. **MeSH Browser**. 2015. Disponível em: <https://www.nlm.nih.gov/mesh/MBrowser.html>. Acessado: 13-11-2015.

MESH. **MeSH Fact Sheet**. 2015. Disponível em: <https://www.nlm.nih.gov/pubs/factsheets/mesh.html>. Acessado: 13-11-2015.

MESH. **MeSH Files**. 2015. Disponível em: <https://www.nlm.nih.gov/mesh/filelist.html>. Acessado: 13-11-2015.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, 2010.

MILLER, K. R.; LEVINE, J. S. **Biology**. [S.l.]: Pearson Prentice Hall, 2010.

MODEL, M. L. **Bioinformatics Programming Using Python: Practical Programming for Biological Data**. [S.l.]: O'Reilly, 2010.

NEKRUTENKO, A.; TAYLOR, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. **Nature Reviews Genetics**, 2013.

NG, S. B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. **Nature**, 2009.

OMIC. **Variant prioritization software tools**. 2015. Disponível em: <http://omictools.com/variant-prioritization-c1654-p1.html>. Acessado: 29-11-2015.

OMIM. **OMIM about**. 2015. Disponível em: <http://www.omim.org/about>. Acessado: 16-11-2015.

OMIM. **OMIM Clinical Synopsis**. 2015. Disponível em: <http://www.omim.org/search/advanced/clinicalSynopsis>. Acessado: 16-11-2015.

PABINGER, S. et al. A survey of tools for variant analysis of next-generation genome sequencing data. **Briefings in Bioinformatics**, 2013.

PERROW, G. **SQL Anywhere Database Backend for Django**. 2015. Disponível em: <https://code.google.com/p/sqlany-django/>. Acessado: 14-11-2015.

PRIYADARSHI, R. **Python support for IBM DB2 and IBM Informix**. 2015. Disponível em: <https://github.com/ibmdb/python-ibmdb>. Acessado: 14-11-2015.

PYRAMID. **Web development with style, your way!** 2015. Disponível em: <http://www.pylonsproject.org/>. Acessado: 18-11-2015.

PYTHON. **DB-API 2.0 interface for SQLite databases**. 2015. Disponível em: <https://docs.python.org/2/library/sqlite3.html>. Acessado: 16-11-2015.

RAILS, R. on. **Web development that doesn't hurt**. 2015. Disponível em: <http://rubyonrails.org/>. Acessado: 18-11-2015.

REED, S. C. **Counseling in medical genetics**. [S.l.]: WB Saunders, 1955.

REENSKAUG, T. M. H. **MVC XEROX PARC 1978-79**. 2015. Disponível em: <https://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html>. Acessado: 14-11-2015.

ROBAINA, M. **Firebird SQL backend for django**. 2015. Disponível em: <https://github.com/maxirobaina/django-firebird>. Acessado: 14-11-2015.

ROBINSON, P. et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. **Genome Research**, 2013.

SAMTOOLS. **The Variant Call Format (VCF) Version 4.1 Specification**. 2015. Disponível em: <http://samtools.github.io/hts-specs/VCFv4.1.pdf>. Acessado: 15-11-2015.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, 1977.

SIFRIM, A. et al. extasy: variant prioritization by genomic data fusion. **Nature Methods**, 2013.

SIMS, D. et al. Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, 2014.

SMEDLEY, D. et al. Walking the interactome for candidate prioritization in exome sequencing studies of mendelian diseases. **Bioinformatics**, 2014.

SQLITE. **SQLite About**. 2015. Disponível em: <https://www.sqlite.org/about.html>. Acessado: 16-11-2015.

STEVENS, T. J.; BOUCHER, W. **Python Programming for Biology**. [S.l.]: Cambridge University Press, 2015.

STITZIEL, N. O.; KIEZUN, A.; SUNYAEV, S. Computational and statistical approaches to analyzing variants identified by exome sequencing. **Genome Biology**, 2011.

TIAN, R.; BASU, M. K.; CAPRIOTTI, E. Contrastrank: a new method for ranking putative cancer driver genes and classification of tumor samples. **Bioinformatics**, 2014.

WANG, Z. et al. The role and challenges of exome sequencing in studies of human diseases. **Frontiers in Genetics**, 2013.

WETTERSTRAND, K. A. **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program**. 2015. Disponível em: <http://www.genome.gov/sequencingcosts>. Acessado: 11-11-2015.

WHITTON, T. **Django by example**. 2012. Disponível em: <http://www.thomaswhitton.com/django-presentation>. Acessado: 22-11-2015.

WU, J.; LI, Y.; JIANG, R. Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. **Public Library of Science Genetics**, 2014.

YII. **The Fast, Secure and Professional PHP Framework**. 2015. Disponível em: <http://www.yiiframework.com/>. Acessado: 18-11-2015.

APÊNDICE A — MESH MEMORANDUM OF UNDERSTANDING

You should read carefully the following Memorandum of Understanding. Downloading MeSH data through this site indicates your acceptance of the following terms and conditions.

MeSH Subject Authority File

MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus. Several MeSH files are available via FTP and have online documentation. MeSH is updated annually and users of this information are encouraged to obtain the new year's data. The data are generally available by mid-November of the preceding year.

Terms and conditions of use

No license is required to obtain the data via FTP. Use of this data is subject to the following restrictions and by obtaining a copy of the data, the user is understood to abide by these conditions:

1. If the use is not personal, (1) the U.S. National Library of Medicine must be identified as the creator, maintainer, and provider of the data; (2) the version of the data must be clearly stated by MeSH year, e.g., 1997 MeSH; and (3) if any modification is made in the content of the file, this must be stated, along with a description of the modifications.
2. Neither the United States Government, nor any of its agencies, contractors, subcontractors or employees makes any warranties, expressed or implied, with respect to data contained in the database, and, furthermore, assumes no legal liability for any party's use, or the results of such use, of any part of the database.
3. You will not assert any proprietary rights to any portion of the database, or represent the database or any part thereof to anyone as other than a United States Government database.
4. The MeSH data carry an international copyright outside the United States, its Territories or Possessions. These terms and conditions are in effect as long as the user retains any of the MeSH data obtained from this site.

Availability

The data are available to all requesters, both within and outside the United States. There is no charge for obtaining the file. Users are required to complete an online registration form before receiving the data.

APÊNDICE B — OMIM USE AGREEMENT

1. Please read this USE AGREEMENT carefully before using this website. This Use Agreement applies to any individual, institution, or organization that uses OMIM.org through its front end, including its mirror sites, the OMIM.org API, or downloads of OMIM data via FTP.
2. OMIM is based on the biomedical literature which is constantly evolving. OMIM data are for research and educational use only. OMIM is not a replacement for the original source or a real-time search of the biomedical literature and is not a substitute for professional medical consultation.
3. By accessing and using OMIM you agree to be bound by all the terms and conditions set forth in this USE AGREEMENT. This USE AGREEMENT shall constitute a legally binding agreement between the user and Johns Hopkins University (hereinafter "JHU"). If you do not wish to be bound by its terms, please refrain from accessing or using OMIM.
4. Users accessing OMIM and OMIM data by implementation of an internet robot are bound by this agreement and the rules in http://omim.org/robots.txt .
5. The rights in and to OMIM (excluding information contained therein obtained from third parties) vest in JHU. JHU holds the copyright and trademark to OMIM and OMIM.org, including the collective data therein and provides access to any individual subject to the terms set forth in this USE AGREEMENT.
6. The license granted hereunder shall become effective on first use and remains in force until terminated as provided in this USE AGREEMENT.
7. Use of OMIM.org is provided free of charge to any individual for personal use, for educational or scholarly use, or for research purposes through the front end of the database. Any individual , commercial and not-for-profit entities and institutions (hereafter called User) wishing to download all or part of OMIM is subject to the terms of this USE AGREEMENT.
8. Users at for-profit or commercial entities who want to download all or part of OMIM must obtain a license by paying applicable licensing fees to and entering into a license agreement with JHU which has the exclusive right to license the access to and use of OMIM to users worldwide. Such license agreement may be in standard form or negotiated between you and JHU. By accessing and using OMIM or related information, you agree to be bound by our standard license agreement and pay any related fees stipulated therein.

In addition, if you access OMIM without a license, you agree to payment of penalties of double the standard license. Requests for information regarding a license for commercial use of the OMIM database may be sent via e-mail to JHTT-Communications@jhmi.edu.

9. Subject to the terms and conditions of this USE AGREEMENT, JHU grants User a non-exclusive, non-transferable and non-sub-licensable license.
10. User may not sell, lease, rent, sublicense, assign, export or transfer in any other manner the OMIM data or the rights of access to or use thereof, or any underlying information, software or other technology, and/or any documents, CDs or any other tangible media representing, embodying or containing any of the foregoing or portion thereof, to any person or entity.
11. No license is granted hereunder to any enhancement or update of or to OMIM.
12. JHU may modify the OMIM data or OMIM.org, including, without limitation, by the removal, reduction or addition of functionality or content; and/or discontinue, temporarily or permanently access to or use of OMIM; and/or change and amend this USE AGREEMENT as provided in clause 28 below, including, without limitation, by adding terms that JHU may be required to "pass through" to User as a result of a separate agreement between JHU and a third party.
13. If OMIM data are downloaded in whole or part from the FTP or API servers and used in a database or analysis software, the data must be refreshed at least weekly. Proper attribution to OMIM and OMIM data must be given.
14. An API key is required to access the OMIM API. This unique KEY will be generated upon registration, must be renewed yearly, and must be included with every request. Johns Hopkins University reserves the right to revoke the key at its discretion.
15. Although steps have been taken to prevent unauthorized alterations or modifications to OMIM and OMIM.org, security mechanisms implemented for OMIM and OMIM.org have inherent limitations. JHU expressly disclaims any warranty that queries to OMIM and/or OMIM.org and other information that Users transmit over the World Wide Web will be protected from third party access. Any loss or damage caused to Users arising out of or in connection with the access to and/or use of OMIM and/or OMIM.org will be borne exclusively by the User, and User agrees that neither JHU nor any of its respective directors, officers, or employees shall have any liability for any such loss or damage.
16. Notwithstanding that the access to and use of OMIM.org is currently free of charge and requires no registration, in consideration for the use, JHU may require User to provide

certain information to JHU. If so requested, the User agrees to: (i) provide true, accurate, current and complete information about User as required (such information, the "User Information"), and (ii) maintain and promptly update the User Information on OMIM.org to keep it true, accurate, current and complete. If JHU requests the User to provide User Information and the User provides such information that is untrue, inaccurate, not current or incomplete, or if JHU has reason to believe the same, then JHU shall have the right to suspend or terminate the license granted hereunder and to refuse to provide the User with any current or future access to and use of OMIM.

17. JHU is entitled to monitor access to and use of OMIM and/or OMIM.org, as long as no personally identifiable information is collected without User's prior consent. JHU may use certain aggregate information related to the use of OMIM and/or OMIM.org, provided that such information will not include personally identifiable information, except as authorized by User. Notwithstanding the foregoing, such personally identifiable information (if collected) may be provided by JHU to third parties in the good faith belief that such action is reasonably necessary to comply with applicable laws, legal process or to enforce this USE AGREEMENT. Furthermore, JHU may use personally identifiable information with respect to commercial users.
18. JHU shall not be obligated to provide User with any support relating to the use of OMIM, except for online documentation contained in OMIM.org.
19. OMIM and/or OMIM.org may provide links to and data from other World Wide Web sites or resources. JHU does not endorse and is not responsible for any data, software or other content available from such sites or resources or their privacy policies. User acknowledges and agrees that JHU shall not be liable, directly or indirectly, for any damage or loss (direct or indirect) relating to User's use of or reliance on such data, software or other content. User shall be solely responsible for obtaining any necessary licenses and/or for compliance with applicable terms of use, as may be required to use data, software or other content from such sites or resources.
20. USER ACKNOWLEDGES THAT OMIM, INCLUDING WITHOUT LIMITATION, THE DATABASE OF INFORMATION CONTAINED THEREIN, IS EXPERIMENTAL AND ACADEMIC IN NATURE, AND IS NOT LICENSED BY THE U.S. FOOD AND DRUG ADMINISTRATION OR ANY OTHER REGULATORY BODY.
21. OMIM IS PROVIDED ON AN "AS IS" AND "AS AVAILABLE" BASIS WITH ALL FAULTS. JHU MAKES NO WARRANTY OR REPRESENTATION OF ANY KIND, EXPRESS OR IMPLIED, WITH RESPECT TO OMIM OR OMIM.ORG INCLUDING,

WITHOUT LIMITATION, ANY WARRANTY OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT OF THIRD PARTY RIGHTS. WITHOUT DEROGATING FROM THE GENERALITY OF THE FOREGOING, JHU DOES NOT WARRANT THAT OMIM WILL MEET USER'S REQUIREMENTS OR THAT USE OF OMIM WILL BE FREE OF INFECTION OR VIRUSES, ERROR-FREE, UNINTERRUPTED, SECURE OR WILL PRODUCE ACCURATE RESULTS. USER SHALL BEAR TOTAL AND EXCLUSIVE RESPONSIBILITY AND RISK FOR THE USE OF OMIM. USER SHALL BE SOLELY RESPONSIBLE FOR ANY RESULTING DAMAGE TO USER'S COMPUTER SYSTEMS OR LOSS OF DATA.

22. USER AGREES THAT JHU AND ITS RESPECTIVE DIRECTORS, OFFICERS AND EMPLOYEES (COLLECTIVELY, THE "INDEMNITEES") SHALL NOT BE LIABLE FOR ANY CLAIMS, DEMANDS, LIABILITIES, COSTS, LOSSES, DAMAGES OR EXPENSES (INCLUDING LEGAL COSTS AND ATTORNEYS' FEES) CAUSED TO OR SUFFERED BY ANY PERSON OR ENTITY (INCLUDING WITHOUT LIMITATION, USER), THAT DIRECTLY OR INDIRECTLY ARISE OUT OF OR RESULT FROM USE OF OMIM BY USER, BREACH OF THIS USE AGREEMENT BY USER OR VIOLATION OF ANY RIGHTS OF ANY THIRD PARTY (ALL OF THE FOREGOING, COLLECTIVELY, "CLAIMS"). WITHOUT DEROGATING FROM THE GENERALITY OF THE FOREGOING, THE INDEMNITEES SHALL NOT BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR PUNITIVE DAMAGES, INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOST DATA OR LOST PROFITS WHETHER ARISING FROM OR RELATING TO TORT (INCLUDING NEGLIGENCE), CONTRACT OR OTHERWISE.
23. User agrees to defend, indemnify and hold the Indemnities harmless from and against any Claims arising from User's use of OMIM.
24. The license granted hereunder will automatically terminate if the User does not comply with this USE AGREEMENT. Either JHU or User may terminate the license granted hereunder and USE AGREEMENT (as between JHU and User) without notice, at any time, for any reason. Upon any termination of the said license and USE AGREEMENT, User's right to use OMIM shall immediately cease. The provisions of clauses 12, 15 and 16 above and clauses 18 and 19 above shall survive the termination of this USE AGREEMENT.
25. User acknowledges and agrees that OMIM, including, without limitation, text, data, com-

pilation, software, sound, photographs, video, graphics or other material contained in or presented to User as part of OMIM and the arrangement thereof, and any JHU software used in connection with OMIM, contain proprietary and confidential information that is protected by applicable intellectual property and other laws. User shall not copy, reproduce, distribute or create derivative works of or otherwise modify OMIM or any part thereof.

26. User agrees not to reverse engineer, recompile, dis-assemble, copy or otherwise attempt to discern the source code of any component of OMIM. JHU reserves the right to terminate any CPU intensive processes. User may not mount an attack against OMIM, attempt to gain root access, or conduct any other activity intended to disrupt OMIM.
27. User may not assign or transfer User's rights under this USE AGREEMENT. JHU may assign all or any of its rights and obligations under this USE AGREEMENT to any third party without User's consent. USE AGREEMENT will inure to the benefit of and be binding upon the successors and assignees of JHU.
28. Changes and amendments to this USE AGREEMENT shall be published on the Website and become effective upon the date of publication. User will not receive a personal notice of such changes and amendments. User agrees to monitor the Website regularly for notices of such changes. By continuing to access and/or use OMIM after such changes and amendments become effective, User shall be deemed to have accepted the modified this USE AGREEMENT.
29. User agrees to access and use OMIM in accordance with all applicable laws and regulations. User agrees not to use OMIM for any illegal or wrongful purposes.
30. If any provision of this USE AGREEMENT is held by a court of competent jurisdiction to be invalid, such provision shall be substituted by a provision which achieves to the greatest extent possible, the same effect as would have been achieved by the invalid provision and all other provisions shall remain in full force and effect.
31. JHU's failure to exercise or enforce any right or provision of this USE AGREEMENT shall not constitute a waiver of such right or provision.
32. This USE AGREEMENT shall be governed by and construed in accordance with the laws of the State of Maryland without regard to or application of choice of law rules or principles. User hereby consents to the exclusive jurisdiction of the competent courts in the state or federal courts of Maryland.

APÊNDICE C — EXEMPLO DE UM REGISTRO DO OMIM

RECORD

***FIELD* NO**

100050

***FIELD* TI**

100050 AARSKOG SYNDROME, AUTOSOMAL DOMINANT

***FIELD* TX**

DESCRIPTION

Aarskog syndrome is characterized by short stature and facial, limb, and genital anomalies. One form of the disorder is X-linked (see 305400), but there is also evidence for autosomal dominant and autosomal recessive (227330) inheritance (summary by Grier et al., 1983).

CLINICAL FEATURES

Grier et al. (1983) reported father and 2 sons with typical Aarskog syndrome, including short stature, hypertelorism, and shawl scrotum. Stretchable skin was present in these patients.

INHERITANCE

Grier et al. (1983) tabulated the findings in 82 previously reported cases of Aarskog syndrome and noted that X-linked recessive inheritance was repeatedly suggested. However, their family had father-to-son transmission, and a family reported by Welch (1974) had affected males in 3 consecutive generations. Grier et al. (1983) suggested autosomal dominant inheritance with strong sex-influence and possibly ascertainment bias resulting from use of the shawl scrotum as a main criterion.

Van de Vooren et al. (1983) studied a large family in which Aarskog syndrome was segregating with variable expression in 3 generations and with male-to-male transmission. Because 3 daughters of affected males had no features of Aarskog syndrome and 2 sons of an affected male had several features of the syndrome, van de Vooren et al. (1983) suggested sex-influenced autosomal dominant inheritance.

***FIELD* RF**

1. Grier, R. E.; Farrington, F. H.; Kendig, R.; Mamunes, P.: Autosomal dominant inheritance of the Aarskog syndrome. Am. J. Med. Genet. 15: 39-46, 1983.

2. van de Vooren, M. J.; Niermeijer, M. F.; Hoogeboom, A. J. M.: The Aarskog syndrome in a large family, suggestive for autosomal dominant inheritance. Clin. Genet. 24: 439-445, 1983.

3. Welch, J. P.: Elucidation of a “new” pleiotropic connective tissue disorder. Birth Defects Orig. Art. Ser. X(10): 138-146, 1974.

***FIELD* CS**

Growth:

Mild to moderate short stature

Head:

Normocephaly

Hair:

Widow's peak

Facies:

Maxillary hypoplasia;

Broad nasal bridge;

Anteverted nostrils;

Long philtrum;

Broad upper lip;

Curved linear dimple below the lower lip

Eyes:

Hypertelorism;

Ptosis;

Down-slanted palpebral fissures;

Ophthalmoplegia;

Strabismus;

Hyperopic astigmatism;

Large cornea

Ears:

Floppy ears;

Lop-ears

Mouth:

Cleft lip/palate

GU:

Shawl scrotum;

Saddle-bag scrotum;

Cryptorchidism

Limbs:

Brachydactyly;

Digital contractures;

Clinodactyly;

Mild syndactyly;

Transverse palmar crease;

Lymphedema of the feet

Joints:

Ligamentous laxity;

Osteochondritis dissecans;

Proximal finger joint hyperextensibility;

Flexed distal finger joints;

Genu recurvatum;

Flat feet

Skin:

Stretchable skin

Spine:

Cervical spine hypermobility;

Odontoid anomaly

Heme:

Macrocytic anemia;

Hemochromatosis

GI:

Hepatomegaly;

Portal cirrhosis;

Imperforate anus;

Rectoperineal fistula

Pulmonary:

Interstitial pulmonary disease

Thorax:

Sternal deformity

Inheritance:

Sex-influenced autosomal dominant form;

also X-linked form

***FIELD* CN**

Nara Sobreira - updated: 4/22/2013

***FIELD* CD**

Victor A. McKusick: 6/4/1986

***FIELD* ED**

carol: 04/24/2013

carol: 4/22/2013

carol: 2/16/2011

alopez: 6/3/1997

mimadm: 3/11/1994

carol: 7/7/1993

supermim: 3/16/1992

supermim: 3/20/1990

ddp: 10/26/1989

marie: 3/25/1988

APÊNDICE D — EXEMPLO DE UM REGISTRO DO MESH

***NEWRECORD**

RECTYPE = D

MH = Tourette Syndrome

AQ = BL CF CI CL CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI

PRINT ENTRY = Gilles de la Tourette's Disease|T048|EPO|EQV|UNK (19XX)|880602|GILLES DE LA TOURETTE DIS|abcdefv

PRINT ENTRY = Tic Disorder, Combined Vocal and Multiple Motor|T048|NON|EQV|NLM (2000)|991103|TIC DIS COMBINED VOCAL MULTIPLE MOTOR|abcdefv

ENTRY = Chronic Motor and Vocal Tic Disorder|T048|NON|EQV|GHR (2014)|130418|abcdef

ENTRY = Combined Multiple Motor and Vocal Tic Disorder|T048|NON|EQV|NLM (2000)|991103|COMBINED MULTIPLE MOTOR VOCAL TIC DIS|abcdefv

ENTRY = Combined Vocal and Multiple Motor Tic Disorder|T048|NON|EQV|NLM (2000)|991103|COMBINED VOCAL MULTIPLE MOTOR TIC DIS|abcdefv

ENTRY = Gilles De La Tourette's Syndrome|T048|NON|EQV|GHR (2014)|130418|abcdef

ENTRY = Gilles de la Tourette Syndrome|T048|EPO|EQV|GHR (2014)|OMIM (2013)|UNK (19XX)|880602|abcdeeef

ENTRY = Multiple Motor and Vocal Tic Disorder, Combined|T048|NON|EQV|NLM (2000)|991103|MULTIPLE MOTOR VOCAL TIC DIS COMBINED|abcdefv

ENTRY = Tourette Disease|T048|EPO|EQV|NLM (2000)|991103|abcdef

ENTRY = Tourette Disorder|T048|EPO|EQV|GHR (2014)|NLM (2000)|OMIM (2013)|991103|TOURETTE DIS|abcdeeefv

ENTRY = Tourette's Disease|T048|EPO|EQV|GHR (2014)|UNK (19XX)|880602|TOURETTES DIS|abcdeefv

ENTRY = Tourette's Disorder|T048|EPO|EQV|UNK (19XX)|880602|abcdef

ENTRY = Tourette's Syndrome|T048|EPO|EQV|UNK (19XX)|880602|abcdef

ENTRY = Syndrome, Tourette's

ENTRY = Tourettes Disease

ENTRY = Tourettes Disorder

ENTRY = Tourettes Syndrome

MN = C10.228.140.079.898

MN = C10.228.662.825.800

MN = C10.574.500.850

MN = C16.320.400.820

MN = F03.550.825.850

FX = Tics

MH_TH = GHR (2014)

MH_TH = NLM (1965)

MH_TH = OMIM (2013)

ST = T048

MS = A neuropsychological disorder related to alterations in DOPAMINE metabolism and neurotransmission involving frontal-subcortical neuronal circuits. Both multiple motor and one or more vocal tics need to be present with TICS occurring many times a day, nearly daily, over a period of more than one year. The onset is before age 18 and the disturbance is not due to direct physiological effects of a substance or a general medical condition. The disturbance causes marked distress or significant impairment in social, occupational, or other important areas of functioning. (From DSM-IV, 1994; Neurol Clin 1997 May;15(2):357-79)

PM = 1989; see GILLES DE LA TOURETTE'S DISEASE 1966-88

HN = 1989(1966)

CATSH = CAT LIST

MR = 20130708

DA = 19990101

DC = 1

DX = 19660101

UI = D005879

APÊNDICE E — ELEMENTOS DE PRIORIZAÇÃO

Consequence (múltipla escolha):

- transcript_ablation
- splice_acceptor_variant
- splice_donor_variant
- stop_gained
- frameshift_variant
- stop_lost
- start_lost
- transcript_amplification
- inframe_insertion
- inframe_deletion
- missense_variant
- protein_altering_variant
- splice_region_variant
- incomplete_terminal_codon_variant
- stop_retained_variant
- synonymous_variant
- coding_sequence_variant
- mature_miRNA_variant
- 5_prime_UTR_variant
- 3_prime_UTR_variant
- non_coding_transcript_exon_variant
- intron_variant
- NMD_transcript_variant
- non_coding_transcript_variant
- upstream_gene_variant
- downstream_gene_variant
- TFBS_ablation
- TFBS_amplification
- TF_binding_site_variant
- regulatory_region_ablation
- regulatory_region_amplification
- feature_elongation
- regulatory_region_variant
- feature_truncation
- intergenic_variant

IMPACT (múltipla escolha):

- HIGH
- MODERATE
- LOW
- MODIFIER

SYMBOL (múltipla escolha): gene ou conjunto de genes

Feature type (escolha simples):

- Transcript
- RegulatoryFeature
- MotifFeature

BIOTYPE (múltipla escolha):

- protein_coding
- pseudogene
- long_noncoding
- short_noncoding
- retained_intron
- processed_transcript

Existing_variation (múltipla escolha):

- COSMIC
- dbSNP

VARIANT_CLASS (múltipla escolha):

- SNV
- genetic_marker
- substitution
- tandem_repeat
- complex_structural_alteration
- copy_number_gain
- copy_number_loss
- copy_number_variation
- duplication
- interchromosomal_breakpoint
- intrachromosomal_breakpoint
- inversion
- mobile_element_insertion
- novel_sequence_insertion
- tandem_duplication
- translocation
- deletion
- indel
- insertion
- sequence_alteration
- probe

SIFT (múltipla escolha):

- probably_damaging
- possibly_damaging
- benign
- unknown
- tolerated
- deleterious

PolyPhen (múltipla escolha):

- probably_damaging
- possibly_damaging
- benign
- unknown
- tolerated
- deleterious

GMAF (escolha simples):

- <= 1 %
- <= 5 %
- > 5 %

CLIN_SIG (múltipla escolha):

- pathogenic
- likely_pathogenic
- benign
- likely_benign

QUAL (valor): qualidade
FILTER (escolha simples): passou ou não por todos os filtros
genotype (escolha simples):

- homozygous
- heterozygous

APÊNDICE F — EXPERIMENTO DE VALIDAÇÃO DA PRIORIZAÇÃO

Tabela F.1 – Resultado do experimento de validação da priorização de variantes

Amostra	*Número de variantes totais*	*Filtros utilizados*									
		Cons	*IMPA*	*Feat*	*BIOT*	*Exis*	*VARI*	*SIFT*	*Poly*	*GMAF*	*CLIN*
26	8	X	X	X	X	X	X	-	-	X	-
27	7	X	X	X	X	-	X	-	-	-	X
28	7	X	X	X	X	X	X	-	X	X	-
29	7	X	X	X	X	X	X	-	-	X	-
30	8	X	X	X	X	X	X	-	-	X	-
31	22	X	X	X	X	X	X	-	-	X	-
32	34	X	X	X	X	-	X	-	-	-	-
33	25	X	X	X	X	X	X	-	-	X	-
34	25	X	X	X	X	-	X	X	X	X	-
35	23	X	X	X	X	X	X	X	X	X	X

APÊNDICE G — DESEMPENHO DO MÓDULO DE PRIORIZAÇÃO

Tabela G.1 – Número de variantes X Tempo de processamento (em segundos)

Tamanho da entrada	*Exp 1*	*Exp 2*	*Exp 3*	*Exp 4*	*Exp 5*	*Média aritimética*
1000	0.480	0.459	0.459	0.467	0.452	0.463
2000	0.897	0.936	0.921	0.961	0.926	0.928
4000	1.903	1.892	1.929	1.941	1.932	1.919
8000	3.887	3.836	3.888	3.825	3.848	3.857
16000	7.775	7.575	7.799	7.638	7.743	7.706
32000	15.418	15.459	15.631	15.212	15.589	15.462
64000	31.890	31.421	31.043	31.708	32.989	31.810
128000	66.402	66.214	68.324	66.391	67.412	66.948
256000	127.552	126.414	126.987	128.234	126.765	127.190

Tabela G.2 – Número de variantes X Volume de memória utilizado (em *megabytes*)

Tamanho da entrada	*Exp 1*	*Exp 2*	*Exp 3*	*Exp 4*	*Exp 5*	*Média aritimética*
1000	33.093	33.093	41.366	33.093	33.093	34.748
2000	33.093	33.093	33.093	41.366	33.093	34.748
4000	49.640	41.366	57.913	49.640	41.366	47.985
8000	66.186	57.913	66.186	74.460	57.913	64.532
16000	107.553	91.007	107.553	99.280	99.280	100.935
32000	173.740	165.467	165.467	157.194	173.740	167.122
64000	289.567	297.841	281.294	264.747	289.567	284.603
128000	562.589	554.315	537.769	554.315	529.495	547.697
256000	1083.811	1058.991	1075.538	1083.811	1075.538	1075.538