**UNIVERSIDADE FEDERAL DO PARÁ**
**INSTITUTO DE CIÊNCIAS BIOLÓGICAS**
**PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR**

HÉLBER GONZALES ALMEIDA PALHETA

# CONSOLIDAÇÃO DE PREDITORES DE PATOGENICIDADES COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL: Aplicações para análise de variantes clínicas.

BELEM/PA
2022

**UNIVERSIDADE FEDERAL DO PARÁ INSTITUTO DE CIÊNCIAS BIOLÓGICAS**
**PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR**

HÉLBER GONZALES ALMEIDA PALHETA

# CONSOLIDAÇÃO DE PREDITORES DE PATOGENICIDADES COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL: Aplicações para análise de variantes clínica.

Tese de doutorado apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal do Pará como requisito para obtenção do grau de Doutor em Genética e Biologia Molecular, área de concentração Bioinformática.

Orientadora: Profª. Drª. Ândrea Kely Campos Ribeiro dos Santos
Coorientador: Prof. Dr. Gilderlanio Santana de Araújo

BELEM/PA
2022

**Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD**
**Sistema de Bibliotecas da Universidade Federal do Pará**
**Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)**

---

P153c PALHETA, HÉLBER GONZALES ALMEIDA.
CONSOLIDAÇÃO DE PREDITORES DE PATOGENICIDADES COM TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL : APLICAÇÕES PARA ANÁLISE DE VARIANTES CLÍNICAS / HÉLBER GONZALES ALMEIDA PALHETA. — 2022.
87 f. : il. color.

Orientador(a): Profª. Dra. Ândrea Kely Campos Ribeiro dos Santos
Coorientação: Profª. Dra. Gilderlanio Santana de Araujo
Tese (Doutorado) - Universidade Federal do Pará, Instituto de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, Belém, 2022.

1. Machine Learn. 2. Random Forest. 3. Meta-prediction. 4. genome-wide. 5. câncer. I. Título.

CDD 576.5

---

# INSTITUIÇÕES PARTICIPANTES E FONTES FINANCIADORAS

## 1. Instituições Participantes

i. Universidade Federal do Pará (UFPA):

- Laboratório de Genética Humana e Médica (LGHM) do Instituto de Ciências Biológicas da Universidade Federal do Pará;

## 2. Fontes Financiadoras

i. Fundação de Amparo à Pesquisa do Estado do Pará (FAPESPA);

ii. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq);

iii. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES);

iv. Rede de Pesquisa em Genômica Populacional Humana/RPGPH (Bio Comp/CAPES)

Dedico este trabalho a todos os pesquisadores, professores, estudantes, profissionais e técnicos que trabalharam incansavelmente no período da pandemia de COVID-19, em especial aos integrantes do LGHM que participaram desse esforço coletivo no enfrentamento da doença.

## AGRADECIMENTOS

A Deus por me conceder a graça de ter pessoas abençoadas ao meu lado que em todas as ocasiões da vida me proporcionaram grandes aprendizados.

A minha amada esposa Ethiene, pelo apoio firme, sacrifício, dedicação e sabedoria. Por ser minha mediadora na manutenção do equilíbrio emocional. Obrigado por ter escolhido compartilhar a vida ao meu lado e por todo amor e cuidado a mim dispensado.

Aos meus filhos Davi (in memorian), Ana Beatriz e Esther, fontes de motivação e renovação, fortalezas da minha alma, que com doçura e divindade me envolvem de bênçãos todos os dias.

Aos meus pais, Carmo (in memoriam) e Glória, por transmitirem sábios ensinamentos, obrigado por toda doação, sofrimentos, superações e alegrias compartilhadas em nossa família.

Aos meus irmãos Carmem, Glaucy, Glauber, Renata, João, Rogério, Mário e Rose, por se fazerem presentes em minha caminhada.

À minha orientadora Ândrea Kely Campos Ribeiro dos Santos por possuir um olhar além da fronteira do conhecimento em busca de novos desafios para o desenvolvimento, não só do aspecto acadêmico e profissional dos seus orientandos, mas especialmente do aspecto humano, agindo sempre de forma humilde, solidária e ética. Obrigado por me acolher na família LGHM e por acreditar em um potencial bruto e conduzi-lo a um estágio superior.

Ao meu Coorientador Prof. Dr. Gilderlanio Santana de Araújo, verdadeiro exemplo de luta e empenho, espelho que me serve de inspiração. Obrigado por toda paciência, entrega, parceria e amizade, pelos conhecimentos compartilhados e trabalhos produzidos. Aprendizados que ressoarão na minha formação acadêmica, profissional e pessoal.

Aos colegas de trabalho do CEPS, principalmente os da coordenadoria de tecnologia de informação, que deram o apoio necessário para que eu pudesse conciliar estudo e trabalho.

Aos meus amigos e colegas do laboratório LGHM que me receberam de braços abertos e são todos fonte de inspiração. Especialmente aos colegas do grupo de Bioinformática que participaram ativamente no desenvolvimento deste trabalho, muito obrigado.

"*Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer.*"

*(Alan Mathison Turing)*

## RESUMO

Há uma complexidade intrínseca no processo de decisão clínica que envolve o diagnóstico genético. As interpretações clínico-genéticas dependem da obtenção de informações mais assertivas sobre as variantes genéticas pelo especialista. O *ClinVar* armazena cerca de 774.000 associações genéticas e registros clínicos, porém um grande conjunto ainda encontra-se com conflito de interpretação (CI) e significado incerto (VUS). O presente estudo tem por objetivo aplicar técnicas de *machine learn* para criar um meta-preditor baseado em Random Forest (AmazonForest) capaz de prever a patogenicidade de variante genética, para apoiar a interpretação de tratamento clínico no nível genômico. Na metodologia, utilizamos dados *ClinVar* anotados com SnpEff/SnpSift(v4.3) para oito preditores de impacto funcional catalogados (FATHMM, SIFT, PolyPhen-2 (HDIV), PolyPhen-2 (HVAR), PROVEAN, MutationAssessor, MutationTaster2 e LRT). Além disso, avaliamos o desempenho de vários algoritmos de aprendizado de representação, como *autoencoders*. Para propor uma melhor estratégia de classificação, aplicamos o AmazonForest na base completa do dbNFSP(4.0) e exploramos os genes envolvidos em dez vias de sinalização relacionadas ao câncer. Da integração do AmazonForest com toda base do dbNSFP, encontramos um conjunto de 3.468.526 variantes altamente suspeitas com patogenicidade (probabilidade de FR >= 0,95). Já nas vias de sinalização relacionada ao câncer (ciclo celular, Hippo, Myc, Notch, NRF2, PI-3-Kinase/Akt, RTK-RAS, sinalização TGF*β*, P53 e *beta*-catenina/WNT, encontramos 935 variantes genéticas, destas 536 variantes genéticas raras com diversidade genética entre as populações continentais. De acordo com dados do COSMIC, 84 variantes raras estão relacionadas a carcinoma, neoplasia linfoide, glioma, melanoma maligno e neoplasia hematopoiética. Em relação ao modelo AmazonForest e a geração de um novo conjunto enriquecido sobre as informações do dbNSFP, fornecemos um potencial recurso computacional para auxiliar em estudos genômicos de doenças complexas como o câncer.



**Palavras-chave:** *machine learn*. *random forest*. tomada de decisão clínica. *meta-prediction*. *genome-wide*. câncer.

## ABSTRACT

Every clinical decision-making process involving genetic diagnosis requires a high degree of complexity. In human stages, in those that depend on the interpretations of a specialist, obtaining more accurate information about the genetic variants becomes fundamental. ClinVar stores about 774,000 genetic associations and clinical records, but one large set still has a conflict of interpretation (CI) and uncertain meaning (VUS). The present study aims to apply machine learning techniques to create a Random Foreset-based meta-predictor (AmazonForest) for genetic variant pathogenicity prediction to support the interpretation of clinical treatment at the genomic level. In the methodology, we used ClinVar data annotated with SnpEff/SnpSift(v4.3) for eight cataloged functional impact predictors (FATHMM, SIFT, PolyPhen-2 (HDIV), PolyPhen-2 (HVAR), PROVEAN, MutationAssessor, MutationTaster2, and LRT). Also, we evaluated the performance of several representation learning algorithms, such as autoencoders, to propose a better classification strategy. We applied AmazonForest to the complete dbNFSP(4.0) database and explore the genes involved in ten cancer-related signaling pathways. From AmazonForest's integration with the entire dbNSFP base, we handled a set of 3,468,526 highly suspected variants with pathogenicity (FR probability >= 0.95). In cancer-related signaling pathways (cell cycle, Hippo, Myc, Notch, NRF2, PI-3Kinase/Akt, RTK-RAS, TGF*beta* signaling, P53, and beta-catenin/WNT we found 935 genetic variants, 536 rare genetic variants with genetic diversity among continental populations. According to COSMIC data, 84 rare variants are related to carcinoma, lymphoid neoplasia, glioma, melanoma malignancy, and hematopoietic neoplasia. Regarding AmazonForest model and the generation of new enriched set variants on top of the dbNSFP, we provide a potential computational resources to assist genomic studies of complex diseases such as cancer.



**Palavras-chave:** machine learn. random forest. clinical decision-making patogenicity. metaprediction. genome-wide. cancer.

**LISTA DE FIGURAS**

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

AUC *Area Under the Curve* (Área sob a Curva)

CNV *Copy Numby Variants* (A variação do número de cópias)

DDBJ *DNA Data Bank of Japan* (Banco de dados de DNA do Japão)

DNA Ácido Desoxirribonucleico

EMBL *European Bioinformatics Institute*

GWAS *Genome Wide Association Studies* (Estudo de Genomica Ampla)

ML *Machine Learn* (Aprendizado de Máquina)

NGS *Nex Generation Sequencing* (Sequenciamento Nova Geração)

NIH *National Institutes of Health* (Instituto Nacional de Saúde-US)

PGH Projeto Genoma Humano

SGBD Sistema de Gerenciamento de Banco de Dados

SNP *Single Nucleotide Polimorphism* (Polimorfirmos Pontuais)

SNV *Single Nucleotide Variant* (Variante de nucleotídeo único)

SVM *Support Vector Machines* (Máquina de Vetores de Suporte)

VCF *Variant Call Format* (Arquivo com Formatacao de Variantes)

## LISTA DE SÍMBOLOS

$\beta$ Letra grega Beta

# Sumário

APÊNDICES

# 1 INTRODUÇÃO

## 1.1 CONTEXTUALIZAÇÃO

O projeto do sequenciamento completo do genoma humano contribuiu expressivamente para novas descobertas estruturais da molécula do DNA, assim como constituiu uma plataforma importante para a observação da variabilidade genética entre as diferentes populações humanas. Seus resultados fundamentaram a base para os estudos funcionais que buscam entender os mecanismos envolvidos no processo de adoecimento associados com as variações genômicas (LANDER *et al.*, 2001; VENTER *et al.*, 2001). Porém, por sua complexidade, o conhecimento estrutural do genoma não traria resultados práticos imediatos, desta forma, os pesquisadores deram prosseguimento ao Projeto Genoma Humano (PGH) voltando sua atenção para as funções exercidas pelos diferentes genes, identificação de proteínas, análises de expressão e estimação da patogenicidade de variantes genéticas (MACARTHUR *et al.*, 2017).

Estruturalmente, os genes vêm sendo estudados muito antes da descoberta da molécula de DNA em 1953 por Watson & Crick, e em 1977 Sanger e colaboradores desenvolveram o primeiro método de sequenciamento que possibilitou a leitura completa das sequências de nucleotídeos que constituíam os genes. Mas foi apenas na era pós-genômica que houve um grande progresso nas tecnologias de sequenciamento de genomas por meio da implementação do sequenciamento de nova geração (NGS), uma abordagem de sequenciamento genômico em larga escala (KULSKI, 2016; GOODWIN; GOODWIN, 2016).

Recentemente, em uma nova análise de comparação entre os dados gerados pelo uso das técnicas de NGS em conjunto com estudo de associação genômica ampla (GWAS), revelou-se a descoberta de quase 2 milhões de pares de base (pb) do genoma humano que não haviam sido reportadas no genoma de referência, além disso, a integração entre diferentes metodologias têm identificado milhares de variantes genéticas associadas com várias doenças

e traços complexos relacionados às doenças (VISSCHER *et al.*, 2017; TIMPSON *et al.*, 2018; MONTINARO; CAPELLI, 2020).

Nesse contexto houve um aumento substancial na geração de informações genômicas que suscitaram a necessidade da utilização de ferramentas bioinformáticas para analisar esses resultados com maior eficácia (SHERMAN; SALZBERG, 2020), dentre elas a análise *in silico*. A análise *in silico* é um conjunto de ferramentas computacionais para inferir a relação genótipo-fenótipo tendo por base a comparação de sequências de nucleotídeos e/ou aminoácidos, sendo essa, considerada uma abordagem mais adequada que as análises experimentais quando se trata de pesquisas com grandes números de dados como ocorre com o uso do NGS (SINGH; MISTRY, 2016).

Assim, vários conjuntos de ferramentas e soluções estão presentes para corroborar em novas descobertas, entretanto, é sempre válido observar as limitações de cada solução utilizada. Uma das limitações encontradas em nosso estudo foi o fato de se constatar que em uma considerada plataforma usada para auxílio em tomada de decisão, o *ClinVar* (LANDRUM *et al.*, 2013), apresenta inúmeras informações pendentes ou inexistentes para determinadas classificações existentes nesta plataforma. Há ainda, os programas de análise *in silico* denominados preditores de patogenicidade que são utilizados para determinar a significância patogênica de novas variantes. Um programa preditor pode ter uma estrutura e abordar métodos diferentes, assim é possível que nos resultados também possam existir diferenças ou interpretações distintas em relação a variante estudada (REIS *et al.*, 2017; BOSIO *et al.*, 2019).

Desse modo, o estudo dos dados do referido projeto proposto pode auxiliar na melhor caracterização de variantes genéticas e suas distribuições entre as populações mundiais, assim como, no entendimento das bases genéticas e moleculares de várias doenças e seu potencial uso na medicina clínica-genômica ou medicina personalizada para o diagnóstico e tratamento das doenças de forma mais precisa (HASPEL *et al.*, 2019).

## 1.2 BIOINFORMÁTICA

Em várias partes do mundo existem, atualmente, diversos estudos sobre o genoma dos seres vivos na terra. Tais estudos têm por finalidade principal descobrir novas relações e funcionamentos do mecanismo da vida.

Hagen (2000), analisando a origem da bioinformática, explica que no início desses estudos, os pioneiros da computação e da biologia não utilizaram o termo "bioinformática" para descreverem seus trabalhos integrados nas áreas de computação e biologia. Eles tinham uma visão clara de como a tecnologia da computação, matemática e biologia molecular poderiam ser combinadas de forma produtiva para responder às questões fundamentais nas ciências da vida, mas não havia, até então, um termo específico para nomeá-la.

No *Oxford English Dictionary* (DICTIONARY, 2019) e *Hogeweg* (2011) a bioinformática é definida como um ramo da ciência relacionada com o fluxo de informações em sistemas biológicos, com o uso de métodos computacionais em várias áreas da biologia e amplamente utilizada em estudos de genéticas e genomas. É uma ciência aplicada que, de forma geral, define uma interface para a união da biologia e ciência da computação, com o uso de conhecimento interdisciplinar da matemática, estatísticas, química, física, linguagens etc. para dar suporte a uma demanda em novas descobertas e auxiliar a ciência da vida (MANDOIU; ZELIKOVSKY, 2008). Com o grande aumento de dados e informações destas ciências surge a necessidade de aplicar técnicas computacionais para compreender e organizar as informações associadas, como exemplo a de macromoléculas biológicas (HASPEL *et al.*, 2019). Podemos dizer que se torna uma ciência preocupada com o uso da computação nas áreas de pesquisas biológicas, como genômica, transcriptômica, proteômica, genética e evolução até a manipulação de dados em alta tecnologia (GOODMAN, 2002).

Um dos principais feitos na área das análises na biologia surgiu com o desenvolvimento das novas plataformas de sequenciamento NGS (*Next Generation Sequencing*), processos de

sequenciamento de DNA que utilizam metodologias diferenciadas com objetivo principal de acelerar as análises dos dados de genômica e diminuir o valor das despesas quanto ao sequenciamento.

Apesar de se diferenciarem consideravelmente entre si, as plataformas NGS baseiam-se no processamento paralelo massivo de fragmentos de DNA. Só para exemplificar, enquanto um sequenciador de eletroforese capilar processa, no máximo, 96 fragmentos por vez, os sequenciadores de nova geração podem ler bilhões de fragmentos ao mesmo tempo (MARDIS, 2008).

Na Figura 1 foi possível verificar a evolução do crescimento em armazenamento da base de dados de sequências brutas em terabases depositadas no NIH (*National Institute of Health*), na execução de análises de alta complexidade com grande quantidade de informações (genômicas, proteômicas e outros dados de diversos seres vivos). Nessas informações obtivemos classificação de genes, genes relacionados à determinada doença, expressão gênicas, estruturas de proteínas, modelos de proteínas interações, ou seja, um enorme potencial de questionamentos advindos dos resultados alcançados. Com isso, compreendemos que aplicar da melhor forma a bioinformática para o entendimento das informações é fundamental para o desenvolvimento da pesquisa translacional, como sua aplicação mais nobre.

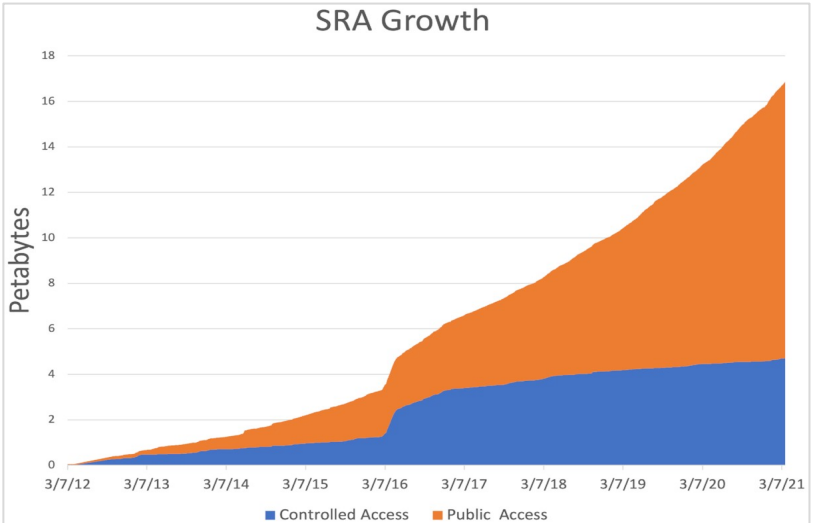



Figura 1: Crescimento exponencial da base de dados SRA (*Sequence Read Archive*), que contém dados de sequências brutas com tecnologia NGS.
Fonte: NIH (*National Institute of Health*).

Sabemos que já existem práticas e linhas de estudos consolidados em bioinformática, contudo, considerando que trata-se de uma ciência que está em constante evolução, nota-se o grande fluxo de conhecimentos ligados à área que ultrapassam seus próprios limites para responder aos novos paradigmas que surgem, gerando sua ampliação, seja na bioinformática, no sentido mais amplo de estudar o processamento de informações, seja na aplicação da biologia de sistemas e Inteligência Artificial (HOGEWEG, 2011).

## 1.3 APRENDIZADO DE MÁQUINA

O "aprendizado de máquina" (ML- *Machine Learn*), é um subcampo da inteligência artificial que surgiu da ideia de criar programas que aprendam determinado comportamento ou padrão automaticamente, baseado em metodologias estatísticas, a partir de exemplos ou observações (MITCHELL, 1997). Segundo o pesquisador, os sistemas de aprendizado de máquinas podem ser aplicados em: i) Representações e armazenamento de conhecimento; ii) Inferências, raciocínio; iii) Aprendizagem, adquirir novos conhecimentos.

O aprendizado de máquina é eficaz quando usado na aquisição automática de descoberta de conhecimento, entretanto, não existe consenso que indique qual técnica apresenta melhor precisão, e /ou melhor desempenho para resolução de problemas de forma generalizada. Há contudo, as que melhor se adaptam a diferentes problemas (ALPAYDIN, 2020).

O ML é o estudo de algoritmos de computador que melhoram automaticamente através da experiência, pois fornece aos sistemas a capacidade de aprender e se aprimorar automaticamente através de dados de treinamento e da descoberta dos processos de execução de tarefas não programadas explicitamente. Isso permite que a máquina desenvolva seu próprio algoritmo, que pode ser usado para prever as saídas em grandes conjuntos de dados (MORAIS, 2020).

Na tecnologia ML temos duas caracterizações importantes quanto a forma de aprendizagem: i) Aprendizagem supervisionada, na qual o especialista humano controla as entradas e saídas de dados desejados e aplica rotulações sobre a precisão das previsões durante a fase de treinamento, e quando finalizado o modelo estará pronto para caracterizar novos dados; ii) Aprendizagem não supervisionada, os algoritmos utilizam uma abordagem interativa, sem a necessidade de um especialista, uma aprendizagem profunda, em geral, usados em processamentos complexos (ALPAYDIN, 2020).

A versatilidade dos modelos classificadores de aprendizado de máquina a torna eficaz na análise de dados e permite sua aplicação em diversas áreas, tais como na análise de dados de saúde, detecção de spam, classificação de tipos de vegetações e decisões financeiras. Basicamente, esses algoritmos de classificação possuem duas etapas, o treinamento, onde o modelo “w” é treinado com exemplos rotulados; e a classificação onde o modelo treinado pode ser executado para gerar uma previsão que indica a qual classe o exemplo pertence (RUSSELL; NORVIG, 2013).

A implementação de estratégias de aprendizado de máquina requer a criação e o treinamento de um modelo. Muitos modelos são bem conhecidos e aplicam diferentes métodos de análises e abordagens de aprendizado de máquina para prever os resultados de saída. Três modelos foram destacados neste estudo: i) *Naive Bayes*; ii) *Random Forest* (Floresta Aleatória); e iii) *Support Vector Machines* (SVM).

### *1.3.1 NAIVE BAYES*

*Naive Bayes* é um classificador probabilístico utilizado em aprendizagem de máquina baseado na utilização do teorema de Thomas Bayes. Esse teorema aplica uma suposição simples, a de que cada variável do seu conjunto de dados assume uma independência condicional entre elas, por esse motivo é também chamada de ingênua na determinação do resultado. Ao aplicar o teorema de Bayes, analisa-se a probabilidade de um evento (A) ocorrer, dado que outro evento ocorreu (B). O método admite que as variáveis envolvidas são independentes, portanto, uma variável não influencia a outra. O teorema de Bayes pode ser calculado:

$$P\,(class/features) = \frac{P\,(class) * P\,(features/class)}{P\,(features)}$$

*Naive Bayes* é um classificador simplista que foi relatado por ter um excelente desempenho em relação ao tempo de execução e precisão, incluindo até mesmo na análise de grandes conjuntos de dados. A utilização da probabilidade bayesiana possibilita que a utilização de um conhecimento prévio de um conjunto de dados e sua lógica, seja aplicado a um conjunto de dados incertos. (BESSIERE et al., 2013; MITCHELL, 2010; RISH et al., 2001).

### *1.3.2 RANDOM FOREST*

*Random Forest* (RF) é um algoritmo de classificação desenvolvido por Leo Breiman (BREIMAN, 2001), que apresenta conhecimentos teóricos avançados nas mais diversas áreas que envolvem a ML. Breiman sugere que a RF trabalhe reduzindo a correlação, mas mantendo uma menor variância. Esse procedimento foi criado com o intuito de contornar as limitações de uma única árvore de decisão. Sinteticamente, RF é um método de classificação por conjunto que combina um conjunto de árvores ajustadas para problemas de classificação ou regressão.

Essencialmente, cada árvore é construída usando um subconjunto de dados por *bootstrap* do conjunto original de amostras, e também por um subconjunto aleatório de recursos (STROBL et al., 2007; LIN; JEON, 2006). Essa amostragem aleatória fornece uma correlação baixa entre as árvores individuais, diminuindo o sobre ajuste (*overfitting*). A decisão final é baseada no cálculo da média da previsão probabilística para cada classe, em vez do voto da maioria (Figura 2).

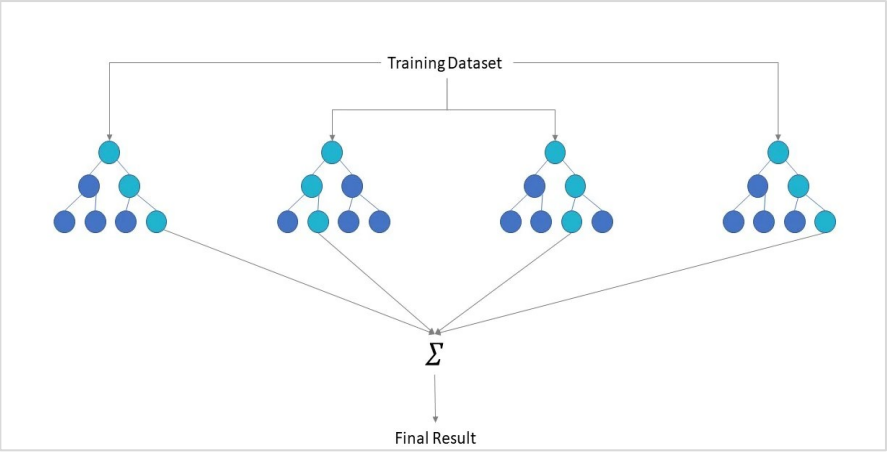



Figura 2: Diagrama do Classificador *Randon Forest*
Fonte: https://www.ibm.com/cloud/learn/random-forest.

O desenvolvimento de aplicações utilizando a técnica de RF ou de outros modelos são variadas, aqui destacamos tipos em que também foram aplicados a técnica de RF, dentre elas

enumeramos apenas algumas como: i) no uso para classificação de tipos de câncer em análises de *microarray*; ii) na utlização para segmentação de imagens; iii) na identificação de interações genéticas; iv) na aplicação de associação do genoma baseado em polimorfismos de nucleotídeo único (CUTLER; CUTLER; STEVENS; 2012).

### *1.3.3 SUPPORT VECTOR MACHINE*

O *Support Vector Machine* (SVM) é uma máquina de aprendizagem proposta em 1992 por Boser, Guyon e Vapnik, que pode ser usada em problemas de regressão ou classificação. Ele usa um procedimento de aprendizagem universal e construtivo baseado na teoria de aprendizagem estatística (CORTES; VAPNIN, 1995).

O SVM propõe buscar em um espaço N-dimensional, onde N é números de características (*features),* para encontrar o hiperplano que melhor separe as classes dos objetos problemas (Figura 1), nesse processo, ele pode também encontrar um mapeamento não linear de dados de alta dimensão, determinando hiperplanos ótimos (XUE; YANG; CHEN, 2009). Este melhor hiperplano descoberto pelo SVM possuirá margens na qual influenciará no grau de eficácia do modelo (BOSER; GUYON; VAPNIK, 1992).

A princípio, o SVM trabalha com conjunto de dados linearmente separáveis pelo hiperplano, porém o SVM leva a separação não linear de dados no espaço de entrada usando funções chamadas de kernels, como mostrado na Figura 3. O SVM também classifica dados que em um primeiro momento não conseguiriam uma classificação direta linear e para solução aplica funções de kernel, que podem ser configurados para qual agrupamento encontram-se os seus dados (LIN; LIN, 2003; HSU *et al.*, 2003) (Figura 4).

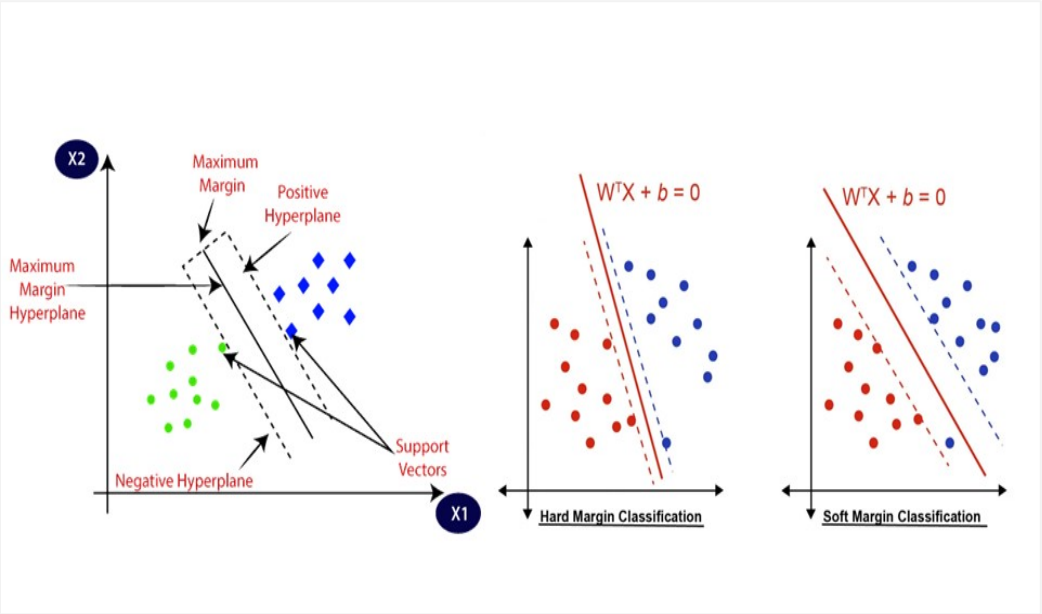



Figura 3: Mapeando dados para um espaço dimensional superior.
Fonte: https://www.analytixlabs.co.in/blog/introduction-support-vector-machine-algorithm, adaptado.

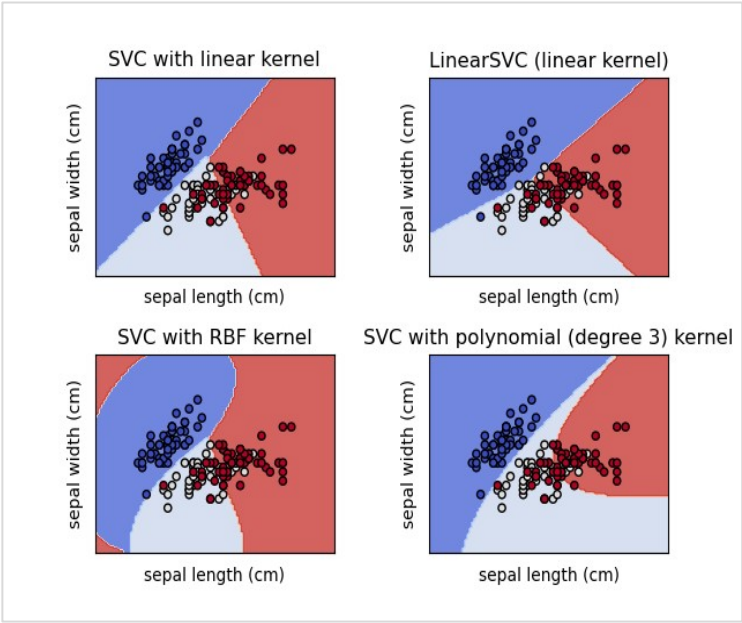



Figura 4: Classificação por SVM, utilizando funções kernel com diferentes tipos para separação dos objetos conforme origem dos dados, agrupamento linear, RBF ou polinomial.
Fonte: https://scikit-learn.org/stable/modules/svm.html.

### *1.3.4 REPRESENTATION LEARNING*

O grande sucesso de um algoritmo de aprendizado de máquina (ML) está intrinsicamente relacionado com a representação (características) de seus dados. Isto porque diferentes representações podem hora misturar ou esconder, para mais ou menos diferentes fatores que poderiam ser mais eficientes nos presentes dados. É claro que o conhecimento de determinado domínio pode ser utilizado para desenvolver e desenhar tais representações. Entretanto, os aprendizados, a princípio com representações de autodescobertas, também podem ser utilizados. A busca por designer de algoritmos de aprendizados de representações estão cada vez mais sendo utilizados (BENGIO; COURVILLE; VINCENT, 2013).

O principal objetivo do aprendizado de representações dos dados (*Representation learning*) é facilitar a extração de informações úteis dos dados para construção de classificadores ou preditores que serão utilizados nos modelos.

Em protótipos probabilísticos a boa representação vai ser aquela que capturar a distribuição posteriori para os fatores mais explicativos para o dado observado. Nas estratégias que podem ser aplicadas podem ter uma abordagem supervisionada ou não supervisionada, aplicação com *Deep Learn* e *auto-econder* (BENGIO; COURVILLE; VINCENT, 2013).

O aprendizado de representação é um dos aspectos importantes para o aprendizado de máquina, na qual descobre-se automaticamente as caraterísticas e os padrões de recursos a serem mais bem utilizados para determinados dados. Desta forma o aprendizado de representação é utilizado no treinamento de algoritmos de aprendizado de máquina para aprender representações úteis, como aquelas que são interpretáveis, incorporam recursos latentes ou podem ser usadas para aprendizado de transferência. Possui utilização para o aprendizado auto supervisionado (SSL, *self-supervised learning*), em campos de visão computacional e NLP (*Natural language processing*) (COATES; NG; LEE, 2011).

### 1.3.5 RECURSOS COMPUTACIONAIS PARA APRENDIZADO DE MÁQUINA

A realidade de tecnologias envolvidas nos conceitos de "aprendizagem de máquina", segundo Hao e Ho (2019), tem tido grande evolução nas últimas décadas e vem sendo relatada nos últimos anos na literatura científica e, por vezes, na imprensa popular. Esse desenvolvimento saudável fez com que mais pacotes de computação que implementam algoritmos de aprendizado de máquina tenham se tornado disponíveis publicamente. Com grande usabilidade, essas implementações amplamente acessíveis ajudam a melhorar e compreender a utilidade e eficácia dos métodos em diferentes contextos, e claro, a expor as limitações através das observações feitas por equipes diferentes em vários domínios de pesquisa ou aplicação.

Hoje encontramos pacotes populares com tais implementações que incluem aqueles com cobertura geral como R (<https://www.r-project.org>) (NWANGANGA; CHAPPLE, 2020), Weka (<https://www.cs.waikato.ac.nz/ml/weka>), *Scikitlearn* (<http://scikit-learn.org/stable>), bibliotecas centradas em redes neurais como *Theano* (<http://deeplearning.net/software/theano>), *Torch* (<http://torch.ch>), *TensorFlow* (<https://www.tensorflow.org>) e *Keras* (<https://keras.io>), além de muitos pacotes focados em aplicativos, como os especializados em visão computacional (OpenCV, <https://opencv.org>), reconhecimento óptico de caracteres (Tesseract, <https://github.com/tesseract-ocr/tesseract>),compreensão da linguagem natural (NLTK, <https://www.nltk.org>; a biblioteca *Stanford* NLP, <https://nlp.stanford.edu>).

O acesso e disponibilização a esses pacotes é cada vez mais eficiente por meio de plataformas de computação em nuvem oferecidas por grandes empresas. Destacamos aqui o pacote *Scikit-learn* que inclui implementações de uma lista abrangente de métodos de aprendizado de máquina sob dados unificados e principalmente por convenções de

procedimentos de modelagem, tornando-o um kit de ferramenta conveniente para aplicações em Aprendizado de Máquina. (INNES *et al.*, 2018; HAO; HO, 2019).

## 1.4 BANCO DE DADOS E PLATAFORMAS GENÔMICAS

Um dos recursos utilizados na área de informática, que também é um dos recursos essenciais para pesquisas e aplicações em bioinformática, são os sistemas de gerenciamento de bancos de dados (SGBD), neles há a união de um conjunto de dados e diversos programas construídos para gerenciar pequenos e grandes volumes de informações. O sistema de gerenciamento dessas informações necessita de mecanismos bem definidos para sua manipulação e segurança das informações armazenadas (SILBERSCHATZ; KORTH; SUDARSHAN, 2011).

Diversos sistemas de banco de dados que são utilizados em bioinformática usam como fontes desde dados primários, com informações biológicas, a repositórios de informações de dados mais completos. Temos bancos com sequência de DNA e proteínas com suas anotações (Ex: Genbank, UniProt); Estruturas de ácidos nucléicos e proteínas (Ex: PDB - Protein Data Bank); Banco de dados específicos para organismos, com o de genomas (Ex: SRA - *Sequence Read Archive*); Banco de dados com redes metabólicas (Ex: KEGG, BioCyc). Dentre outros, com informações secundárias obtidas utilizando-se dados primários, além dos bancos de dados para referências bibliográficas e sistemas de apoio para funcionamento web.

Uma dessas plataformas é mantida pelo NCBI -*National Center for Biotechnology Information* - (Centro Nacional de Informação Biotecnológica), que faz parte do Instituto Nacional de Saúde dos Estados Unidos. O instituto foi criado em 1988 tendo por um dos

objetivos desenvolver sistemas de informações para a biologia molecular (SAYERS *et al.*, 2012). O NCBI é formado por uma série de bancos de dados relacionais, textuais, big data, a exemplo do *GenBank* (BENSON *et al.*, 2012). Há também o *ClinVar* Landrum *et al.* (2013), banco importante por trazer informações clínicas de doenças. É formado com a colaboração de outros bancos de dados como: DDBJ *Bank of Japan* e EMBL-Bank *European Bioinformatics Institute*. Além de prover inúmeros recursos para análises bioinformáticas.

## 1.5 CLINVAR

O *ClinVar* é um repositório de acesso público, e um dos disponibilizados pelo NCBI, possui informações das relações entre variantes gênicas humanas e seus fenótipos que passou por um processo de evidenciação. Estas informações são enriquecidas por um padrão de submissão de informações, como: descrições clínicas de pacientes, seus históricos, suas devidas relações entre estas variantes, interpretações e conclusões (LANDRUM *et al.*, 2013).

No *ClinVar* cada variante depositada será classificada de acordo com ACMG-AMP (*American College of Medical Genetics and Genomics and the Association for Molecular Pathology*): "*benign, likely benign, variant of uncertain significance, likely pathogenic, or pathogenic e interpretation conflict*".

A composição do *ClinVar* está relacionada a outros dois importantes bancos de dados: O dbSNP que mantém informações de SNP ("*Single nucleotide polymorphism* – Polimorfismos pontuais"), de pequenas escalas de deleções ou inserções e repetições de microssatélites (PHAN, 2022); o dbVar, um banco de dados de larga escala com as variantes genômicas, o dbVar o qual complementa também o dbSNP, mantendo CNV (*copy number variants*), inserções, deleções, inversões e translocações (SAYERS *et al.*, 2019). Para nosso estudo

utilizamos o arquivo em formato VCF (obtido em 2 de outubro de 2020, https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/).

## 1.6 VARIANTES GENÉTICAS

O mapeamento completo do "genoma humano referência" abriu caminho para os avanços da genômica e a investigação de variantes no DNA, bem como, nos estudos de seus impactos em traços complexos, doenças e nos processos fundamentais que moldam a variação genética humana que além das variantes, engloba a recombinação e a seleção natural (CONSORTIUM *et al.*, 2015).

Conceitualmente, as variantes genéticas são mudanças em regiões específicas na sequência de DNA, que podem ou não, alterar a estrutura e função do gene, assim como sua proteína correspondente. Essas variantes podem resultar de um erro na replicação do DNA durante a divisão celular, exposição à radiação, mutagênicos químicos ou doenças infecciosas. As variantes identificadas no genoma incluem:

i) Polimorfismos de nucleotídeo único (SNPs), são as alterações mais comuns do genoma humano e consistem na substituição de um nucleotídeo por outro (CONSORTIUM *et al.*, 2015);

ii) Inserções/deleções (INDELs): é a segunda alteração mais comum do genoma, consiste na adição ou remoção de um ou mais nucleotídeos em uma sequência de DNA. Esse tipo de variante, quando em região codificadora, modifica a sequência de nucleotídeos seguintes alterando, assim, os códons e os aminoácidos correspondentes de forma que a proteína resultante é diferente em estrutura e função da proteína original. Por essa razão essa

variante é designada *frameshift* ou mutação de mudança de matriz de leitura (FEUK; CARSON; SCHERER, 2006).

iii) Variantes estruturais (SVs): envolvem grandes inserções ou deleções de pb (mais de 50 pb), variações no número de cópias como deleções e duplicações de cromossomos e outros rearranjos cromossômicos como inversões e translocações (CONSORTIUM *et al.*, 2015; SUDMANT *et al.*, 2015). Apesar de serem as variantes menos comuns e menos estudadas contribuem substancialmente para a diversidade genética, sendo de grande importância evolutiva e biomédica (ALMARRI *et al.*, 2020).

As mutações pontuais que englobam os SNPs de um único nucleotídeo, podem ser mutações sinônimas ou não sinônimas. As mutações sinônimas ocorrem geralmente em regiões codificantes, porém sem alterar a sequência de aminoácidos traduzidos, por isso, são também denominadas de mutações silenciosas (HARTL; CLARK, 2010). Em contrapartida, as mutações não sinônimas compreendem a substituição de um nucleotídeo em regiões codificantes, ocorrendo a alteração do aminoácido produzido pelo códon, e como mencionado anteriormente, isto pode alterar a estrutura da cadeia polipeptídica e resultar na não funcionalidade da proteína.

As mutações não sinônimas podem ainda ser do tipo *missense* (sentido trocado) ou *nonsense* (sem sentido). Na mutação de sentido trocado há a mudança de um códon que especifica um aminoácido por outro que codifica um aminoácido diferente, dessa forma, muitas dessas variantes possuem relevância clínica (RUWALD *et al.*, 2016).

As mutações de sentido trocado podem ser classificadas como conservativas ou não conservativas. As mutações são conservativas quando um aminoácido selvagem é substituído por outro homólogo, ou seja, com propriedades químicas similares, por exemplo, a substituição da isoleucina por leucina ou do aspartato por glutamato (JONSON; PETERSEN, 2001). As mutações são classificadas como não conservativas quando a substituição do aminoácido

ocorre por outro com propriedades diferentes em relação ao tipo selvagem, assim, este tipo de mutação possui maior impacto na estrutura e funcionamento da proteína do que as mutações conservativas (BATMANIAN; RIDGE; WORRALL, 2011).

As mutações sem sentido é a substituição de um único pb que leva ao surgimento de um '*stop codon*' ou códon de parada prematuro e ao interrompimento da tradução (RUWALD et al., 2016), resultando na produção de uma proteína mais curta do que a selvagem que pode ser não funcional ou de funcionamento anormal (RAUSELL et al, 2014).

A maior parte das pesquisas sobre a base genética de doenças tem se concentrado em estudar variantes nas regiões codificantes de proteínas, no entanto é importante enfatizar que uma grande proporção da região não codificante possui funções no genoma, como no empacotamento do DNA, na organização da cromatina no núcleo e na regulação da expressão gênica (FRENCH; EDWARDS, 2020). Dessa forma, variantes nas regiões não codificantes podem estar relacionadas com doenças humanas (FRENCH; EDWARDS, 2020).

## 1.7 JUSTIFICATIVA E RELEVÂNCIA

O *ClinVar* é uma das ferramentas mais utilizadas no apoio de análise de variantes clínico-patológicas, entretanto, como cada adição de informações sobre novas variantes passam por um processo que abrange várias etapas e demanda determinado tempo, temos uma lacuna de informação que pode ficar defasada, sem conclusão ou incompleta.

Segundo Houdayer *et al.* (2012), hoje já existem sistemas baseados em algoritmos utilizando técnicas de inteligência artificial que são utilizados para ajudar em uma melhor tomada de decisão médica ou mesmo fazendo parte de algum protocolo clínico.

Os estudos com sequenciamento de próxima geração (NGS) e estudos de associação ampla do genoma (GWAS) nos permitem explorar as variantes genéticas e seus efeitos patogênicos em determinadas doenças como câncer (FREEDMAN *et al.*, 2011), doença de Alzheimer (NAJ *et al.*, 2014; LOGUE *et al.*, 2011), diabetes tipo 2 (XUE *et al.*, 2018).

No campo do estudo *in silico* de variantes existem alguns preditores de patogenicidade, ferramentas que usam um conjunto de instruções para determinar a patogenicidade de uma variante. Nesse contexto identificamos no *ClinVar* uma grande quantidade de variantes com conclusões incompletas e percebemos a necessidade do desenvolvimento de uma plataforma que agregasse mais informações a esses dados.

Daí surgiu a necessidade de aplicar técnicas de aprendizado de máquinas para unir e analisar resultados dos preditores *in silico*, com intuito de melhorar a informação do conjunto de preditores com maior eficiência para ajudar na caracterização do impacto que essas variantes possuem na determinação de doenças complexas.

# 2 OBJETIVO

## 2.1 OBJETIVO GERAL

Criar um meta-preditor *in silico* para reclassificação de variantes clínicas utilizando informações genômicas da plataforma *ClinVar* adicionadas com informações de anotações funcionais dadas pelos preditores de patogenicidade que serão utilizados para o treinamento de modelos de aprendizado de máquina. Consequentemente, pretende-se enriquecer no resultado final com a probabilidade desta variante ser classificada como patogênica ou não, contribuindo para uma melhor informação na tomada de decisão e diagnóstico clínico.

## 2.2 OBJETIVOS ESPECÍFICOS

- Analisar comparativamente variantes e/ou polimorfismos genéticos do *Clinvar* para treinamento e reclassificação dos VUS / CI;
- Investigar aplicação de transformações de dados dos valores dos preditores para treinamento de modelos de aprendizagem de máquina;
- Investigar aplicações de modelos de “Aprendizado de Máquina” em dados resultantes de preditores, indicando sua patogenicidade;
- Elaborar um novo modelo de meta-predição.
- Elaborar ferramenta que seja facilitadora para execução, análises e utilização do modelo, e apoio a tomada de decisão em diagnóstico clínico.

# 3 CAPITULO I- AMAZONFOREST: IN-SILICO PATHOGENIC METAPREDICTION OF PATHOGENICS VARIANTS

Publicado na revista Biology, (IF 5.168)

*biology*

MDPI

*Article*

# AmazonForest: In Silico Metaprediction of Pathogenic Variants

**Helber Gonzales Almeida Palheta [1], Wanderson Gonçalves Gonçalves [1,2], Leonardo Miranda Brito [1], Arthur Ribeiro dos Santos [1], Marlon dos Reis Matsumoto [1], Ândrea Ribeiro-dos-Santos [1,2,†] and Gilderlanio Santana de Araújo [1,*,†]**

[1] Laboratory of Human and Medical Genetics, Graduate Program of Genetics and Molecular Biology, Institute of Biological Sciences, Federal University of Pará, Belém 66075-110, Brazil; hpalheta@gmail.com (H.G.A.P.); wandersongegoncalves@gmail.com (W.G.G.); lb9458@gmail.com (L.M.B.); arthurrdsantos@outlook.com (A.R.d.S.); marlonmatsumotosdb@gmail.com (M.d.R.M.); akelyufpa@gmail.com (Â.R.-d.-S.)
[2] Research Center on Oncology, Graduate Program of Oncology and Medical Science, Federal University of Pará, Belém 66073-000, Brazil
* Correspondence: gilderlanio@gmail.com
† These authors contributed equally to this work.

**Simple Summary:** ClinVar is a valuable platform that stores a large set of relevant genetic associations with complex phenotypes. However, the functional impact of a partial set of such associations remains misinterpreted, due to the presence of variants with uncertain significance or with conflicting pathogenicity interpretations. To fill this gap, we present AmazonForest: a metaprediction model based on Random Forest for pathogenicity prediction. AmazonForest was used to reclassify a set of ~101,000 variants that were predicted as having high pathogenic probability. AmazonForest is available as a web tool with a simple web interface, and also as an R object for pathogenicity predictions.

**Abstract:** ClinVar is a web platform that stores ~789,000 genetic associations with complex diseases. A partial set of these cataloged genetic associations has challenged clinicians and geneticists, often leading to conflicting interpretations or uncertain clinical impact significance. In this study, we addressed the (re)classification of genetic variants by AmazonForest, which is a random-forest-based pathogenicity metaprediction model that works by combining functional impact data from eight prediction tools. We evaluated the performance of representation learning algorithms such as autoencoders to propose a better strategy. All metaprediction models were trained with ClinVar data, and genetic variants were annotated with eight functional impact predictors cataloged with SnpEff/SnpSift. AmazonForest implements the best random forest model with a one hot data-encoding strategy, which shows an Area Under ROC Curve of ≥0.93. AmazonForest was employed for pathogenicity prediction of a set of ~101,000 genetic variants of uncertain significance or conflict of interpretation. Our findings revealed ~24,000 variants with high pathogenic probability ($RF_{prob} \geq 0.9$). In addition, we show results for Alzheimer's Disease as a demonstration of its application in clinical interpretation of genetic variants in complex diseases. Lastly, AmazonForest is available as a web tool and R object that can be loaded to perform pathogenicity predictions.

**Keywords:** metaprediction; encoding data; random forest; representation learning; genetic variants; clinical impact; functional impact



**Citation:** Palheta, H.G.A.; Gonçalves, W.G.; Brito, L.M.; dos Santos, A.R.; dos Reis Matsumoto, M.; Ribeiro-dos-Santos, Â.; de Araújo, G.S. AmazonForest: In Silico Metaprediction of Pathogenic Variants. *Biology* **2022**, *11*, 538. https://doi.org/10.3390/biology11040538

Academic Editor: Wojciech Makalowski

Received: 26 January 2022
Accepted: 2 March 2022
Published: 31 March 2022



## 1. Introduction

Next-generation sequencing (NGS) methods have allowed whole-genome analyses for humans and other species. Genome-wide association studies (GWAS) and candidate gene studies have produced a large volume of genetic associations between single-nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) with complex diseases. Most of these associations show variable effects and genetic diversity among populations [1,2].

Variants with highly pathogenic effects are responsible for developing several types of cancer [3], Type 2 diabetes [4], and Alzheimer's disease [5,6]. Understanding the biological role and impact of these variants on clinical and personalized levels is a complex task.

ClinVar is an online database that stores around 789,000 curated entries that show associations between phenotypes and genetic variants (SNPs or INDELs) and their clinical relevance (classified as either benign or pathogenic) [7]. ClinVar has improved our understanding of the functional role of genetic variants as research increasingly focuses on precision medicine [8]. However, many genetic variants are functionally misinterpreted and continue to have conflicting interpretations (CI) or uncertain significance (VUS).

Distinct machine learning (ML) metaprediction models have been proposed for pathogenicity prediction of genetic variants, aiming to combine the strengths of multiple pathogenicity prediction programs. Each metaprediction model has been suggested for the analysis of a single variant class (synonymous or nonsynonymous variants) [9–13], and most metapredictors were used for the pathogenicity prediction of VUS and CI variants [9,10,12,13]. Interestingly, most recently proposed metapredictors are decision tree-based or an ensemble of decision trees, which constitute models with clear interpretations. Ensemble-based methods, such as Random Forest (RF), are promising for pathogenicity prediction of coding and noncoding variants [9–11,13]. However, these models have shown differences regarding data-training methods, specifically on data heterogeneity and on the number of features used to train and test each classification model.

Thus, we implemented AmazonForest, a pathogenicity metapredictor based on Random Forest and functional impact data for high confidence pathogenicity interpretation. AmazonForest is the main contribution of this work. In addition, we employed the AmazonForest model to reclassify 100,805 genetic variants, and make available a dataset of ~24,000 genetic variants with high pathogenic probability ($RF_{prob} \geq 0.9$). The resulting dataset sums as a large collection of annotated potentially disease-causing variants that may aid in the investigation and modeling of diseases.

## 2. Materials and Methods

### *2.1. Fetch ClinVar .vcf File*

The first step consists of fetching genome-wide and clinical data from ClinVar, which is stored in .vcf files. The .vcf file is available at https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/, accessed on 2 February 2021. The dataset showed 789,419 genetic variants. Each variant is classified according to the ACMG-AMP [14] with labels that correspond to the following categories: benign, likely benign, variant of uncertain significance, likely pathogenic, pathogenic, or conflict of interpretation.

### *2.2. Functional Impact Variant Annotation by Single Predictors*

SnpEff and SnpSift (v.4.3) configured with dbNSFP4.0 were used for functional annotation of variants stored in ClinVar .vcf files. Therefore, our metapredictor was built based on categorical data extracted from eight predictors: FATHMM, SIFT, PolyPhen-2 HVAR, PolyPhen-2 HDIV, PROVEAN, MutationAssessor, MutationTaster2, and LRT. Each predictor is independent and based on distinct genomic approaches such as sequence characteristics, conservation, and amino acid changes. All predictors are described in detail as follows:

- FATHMM predicts the functional effects of coding and noncoding variants. This predictor combines wild-type and mutated sequences in a hidden Markov model, which identifies mutations in peptide chains, showing the alignment of homologous sequences and conserved protein domains [15];
- SIFT (Sorting Intolerant From Tolerant) is a prediction tool that codes an algorithm for amino acid substitution analyses. It assumes that important positions in a protein sequence have been conserved throughout evolution, and therefore substitutions at these positions may affect protein function. The algorithm sorts changes in a polypeptide chain as tolerant or intolerant according to its evolutionary conservation [16];

- Polyphen-2 (Polymorphism Phenotyping v2) predicts the impact of amino acid substitutions on structural stability, physical interactions, and human protein function. The probability of a mutation being pathogenic depends on the extraction of sequence annotations, structural attributes, and conservation profiles in protein-coding regions [17];
- PROVEAN (Protein Variation Effect Analyzer) is a predictor that provides a generalized approach to predict the functional effects on variations in a peptide chain. These effects include SNPs, INDELs, or multiple amino acid substitutions. Prediction is performed by employing a mutation database obtained from UniProtKB/Swiss-Prot and other experimental data previously generated from mutagenesis experiments [18];
- MutationAssessor predicts the functional impact of amino acid substitutions on proteins using the evolutionary conservation of the affected amino acid in protein counterparts. Multiple Sequence Alignment is used to reflect functional specificity, represent the functional impact of a missense variant, and generate conservation scores. Variants with higher scores are more likely to be pathogenic [19];
- MutationTaster2 predicts functional changes in DNA sequences. It is designed to predict consequences based on amino acid substitutions, and intronic substitutions such as synonymous changes, short insertion or exclusion mutations, and variants that cover the limits of introns and exons [20];
- Likelihood Ratio Test (LRT) is a metric that evaluates the proportion of synonymous and nonsynonymous mutations in protein-coding regions. The altered proportion of mutations means that a negative selection process occurred over that region during evolution, which consequently modifies codons in peptide chains [21].

### *2.3. Encoding Genome-Wide Training and Test Dataset*

After functional annotation, we preprocessed ClinVar data according to ACMG-AMP pathogenicity labels. In this step, we grouped these classes into labels: (a) benign/likely benign into benign; (b) pathogenic/likely pathogenic into pathogenic; (c) variant of uncertain significance and variant with a conflict of interpretation remained with the same label.

A second round of data preprocessing was performed for filtering ClinVar data to avoid variants with missing data. The training/test dataset comprised only variants that were classified by the eight aforementioned single predictors of functional impact. Following functional annotation, the ClinVar dataset was preprocessed using in-house scripts for data extraction and encoding methods. For this study, we investigate data-encoding strategies and representation-learning strategies:

- Label encoding is an approach that assigns numerical values from 0 to the number of classes $-1$ to each of the categorical values in a dataset. For example, if the column with categorical values contains five classes, then the label encoding assigns numerical values between 0 and 4;
- One hot encoding transforms categorical variables using a dummy strategy. Each variable category is transformed into a binary column. For example, given a dataset with two categories, the one hot encoder creates two new columns to store binary values, 0 or 1;
- Multiple Correspondence Analysis (MCA) is a statistical method that handles categorical variables for dimensionality reduction. MCA is an extension of simple correspondence analysis and a generalization of principal component analysis, which is appropriate for quantitative data [22]. The MCA is used to create a low-dimensional space for samples and predictor points based on a contingency table, and the dimensions are retained as eigenvalues;
- Autoencoders are unsupervised learning algorithms that aim to obtain a data representation by reconstructing the input data at the output [23]. In this study, artificial neural networks were implemented to learn representations of the ClinVar data. We used an autoencoder similar to a multilayer perceptron (MLP), with an input layer, a hidden layer with 10, 20, and 30 neurons, and an output layer with the same num-

ber of predictors. Rectifier (ReLu), Rectifier with Dropout, and Hyperbolic Tangent Function (Tanh) were used as neuron activation functions. Dropout is commonly used to reduce overfitting and can improve the results of a classifier. The function of this regularization layer is to turn off a portion of the neurons, forcing the network to readjust the weights and preventing the network from memorizing the training data [24].

### *2.4. Fine-Tuning of Random Forest*

RF is a machine learning method created to avoid the limitations of single predictors, being an ensemble method that combines decision trees for classification or regression problems [25]. Essentially, each tree handles a subset of bootstrapped data from the original set of samples, as well a random subset of predictors [26]. This random sampling raises a low correlation between individual decision trees, which avoids overfitting. The prediction probability for each class is used to reach a final decision and take a majority vote.

We performed a grid-search strategy for fine-tuning RF models taking as input the categorical data, one hot encoded data, and representation-learned data extracted from MCA and autoencoders. The grid search strategy targets two RF parameters: (a) the number of trees in the forest model, that ranges from 50 to 1000 decision trees, and (b) the number of bootstrapped predictors (p), that was set to 2, $\sqrt{p}$, p/2, p. The parameter values were chosen based on experiments from [5,27]. Thus, we defined three experiments, as follows:

1. RF were trained with categorical data and one hot encoder;
2. RF were trained with two extracted MCA dimensions;
3. RF were trained with two dimensions from autoencoders based on three diferent activation functions: rectifier, rectifier with dropout and tahn. Moreover, we range the number of epochs and hidden neurons on autoencoders, which were set for 10, 20, and 30 for both parameters.

For model evaluation, we considered the Area Under Curve (AUC) and the out-of-bag error (OOBE), a strategy similar to cross-validation [28]. AUC is derived from Receiver Operating Curves and represents the degree of class separability, in which values close to 1 represent high-grade model performance. All models were implemented using R base and *randomForest* (v.4.6-14) and *h20* (v.3.34.0.3) libraries.

### *2.5. AmazonForest: Web Platform for Variant Classification*

We developed the online version of AmazonForest to improve user experience on pathogenicity prediction. AmazonForest was implemented as an online platform that performs our best metaprediction model to predict the pathogenicity of VUS, CI, and new genetic variants. AmazonForest is available at https://www2.lghm.ufpa.br/amazonforest, accessed on 6 February 2022. The platform is divided into two components:

- The first is the user interface component. AmazonForest was developed as a web tool with an interface that allows performing pathogenicity prediction of SNPs or INDELs with in silico analyses employing the best metapredictor model. The simple web interface enables the user to predict pathogenicity in two ways. First, by providing genomic or dbSNP information (chromosome, chromosome position, or rsID) and second, by allowing the combination of predictor results to query pathogenicity status. The web component was developed using Python3.6 [29], Javascript (https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference), HTML5 (https://developer.mozilla.org/pt-BR/docs/Web/HTML/HTML5), and using frameworks such as Flask (v.2) (https://palletsprojects.com/p/flask/), scikit-learn [30], Pandas (v.1.1.5) [31], Numpy (V.1.19.5) [32]. All packages was acessed on 2 February 2021.
- The second is a model administrator component to assess the evolution and performance of the model. This model component enables the reproducibility of up-to-date data.

## 3. Results

### *3.1. Training and Test Data Records*

The filtering strategy for ClinVar's database resulted in a slightly unbalanced training dataset without missing data, and more benign variants were cataloged than pathogenic variants. A view of the set of variants in this process is shown in Table 1, which highlights the original number of cataloged variants in the ClinVar database, the distribution of variants by class for the training/test dataset, and the reclassified dataset. Furthermore, data preprocessing showed a significant decrease in the number of genetic variants with functional annotation for each of the eight predictors. The distribution of categorical data was drawn in Figure 1, which highlights the challenge and complexity of interpreting the functional impact of variants. Additionally, we established the number of epochs and hidden neurons on autoencoders, which were set for 10, 20, and 30 for both parameters.

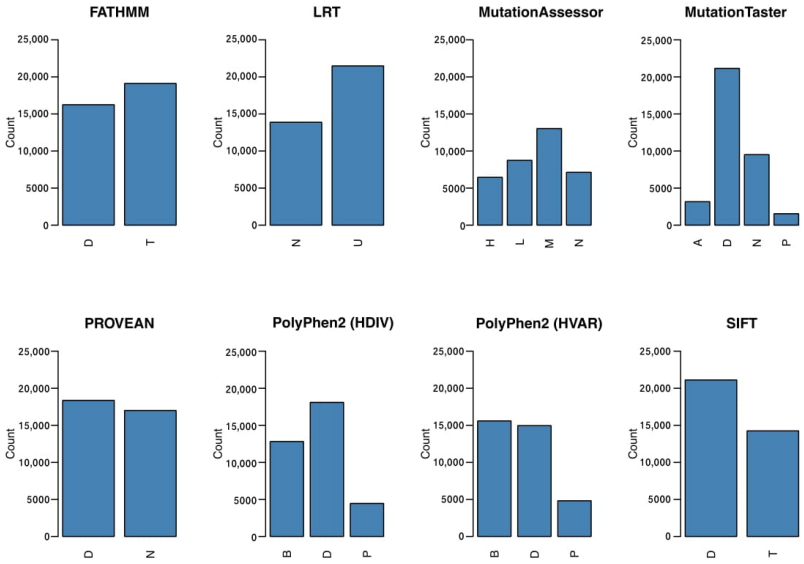



**Figure 1.** Distribution of variants by functional impact prediction for the eight predictors described in Section 2.2. Each functionial predictor provides their own type of classification. Deleterious (D) and Tolerated (T) for FATHMM; neutral (N) or unknown (U) for LRT; high (H), medium (M), low (L), or neutral for MutationAssessor; disease-causing, automatic prediction (A), disease-causing (D), probably harmless automatic prediction (N), and known to be harmless (P) for MutationTaster; deleterious (D) an neutral (N) for PROVEAN; probably damaging (D), possibly damaging (P) and benign for Polyphen; and finally, deleterious (D) and tolerated (T) for SIFT.

**Table 1.** Distribution of genetic variants by functional impact in ClinVar original dataset. The training and test dataset is composed of biological annotated variants for the eight functional predictors described in Section 2.2.

| Category of Genetic Variants in CinVar | Original Dataset | Training Dataset | Reclassification Dataset |
|---|---|---|---|
| Benign | 266,145 | 18,891 | - |
| Pathogenic | 130,739 | 16,471 | - |
| With conflit of interpretation | 42,609 | - | 7193 |
| With uncertain significance | 349,926 | - | 93,612 |

### 3.2. Fine-Tuning and Selection of Metaprediction Model

RF training yielded 144 accurate models for variant pathogenicity prediction. All fine-tuning experiment results were drawn in Figure 2. In the experiments, AUC ranged from 0.88–0.91 using label encoding data, 0.88–0.92 when one hot encoded data were employed, 0.88–0.89, and 0.81–0.89 for representation learned data extracted from MCA and deep autoencoders, respectively. The best RF model reached higher AUC value with 1000 trees and two bootstrapped predictors under training. This model was trained with one hot encoded data and showed an AUC of 0.93 and an OOBE of 14.1% (see Figure 2A). For this model, feature importance analysis by Gini impurity (GI) identified PROVEAN and MutaTaster as the most influential features (GI $> 0.2$). In decreasing order (GI $\leq 0.1$) of importance, GI identified PolyPhen_Hvar, SIFT, PolyPhen2_HDIV, FATHMM, LRT_pred, and MutaAss (see Supplementary Figure S1).

Extraction of representation-learned data from MCA and autoencoder models did not reach higher AUC combined with RF, but are satisfactory models. Compared with label encoding and one hot encoding, the RF model showed the lowest AUC when trained with representation-learned data from MCA or autoencoder data. RF trained with autoencoder data extracted from deep learning models, with Rectifier and Tanh activation functions performed similarly. Most of the AUC for these experiments overlapped (see Figure 2). In contrast, AUC is lower for all the experiments using RF models with autoencoder data from deep learning models trained with rectifiers with dropout. Additionally, we observed the lowest AUC for autoencoders set with rectifier with dropout, higher values in the number of hidden neurons, and trained with a higher number of epochs (see Figure 2C).

In addition to the aforementioned model comparisons, we compared RF, with Naive Bayes (NB), and Support Vector Machine, which showed satisfactory prediction performance, AUC $> 0.9$ (see Supplementary Table S2 and Figure S2). All models were evaluated by performing 10-fold cross-validation. The SVM model trained with linear kernel showed similar results to RF (AUC = 0.93, $+/-0.01$). Based on this evaluation analysis and characteristics of RF and SVM, we chose RF for further analysis, given that SVM performes better on noncategorical data, has a costly computational complexity and high training time for large databases.

### 3.3. Reclassification of VUS and CI Variants

The best RF model was applied to classify 100,805 genetic variants labeled as variants of uncertain significance or conflict of interest. As a result, 32,398 (32.14%) VUS and 2282 (2.26%) CI variants were labeled as pathogenic variants. Out of this last set, we identified a set of 24,428 genetic variants with high-probability of pathogenicity according to RF predictions ($RF_{prob} \geq 0.9$, see Figure 3A). These variants were distributed throughout 1019 gene regions. Reactome pathway analysis was performed for those genes, which revealed a set of 24 enriched pathways (Supplementary Table S1). The enriched pathways are associated with many important cell functions, such as metabolic processes, cell growth and division, extracellular matrix organization and degradation, muscle contraction, and cardiac conduction. Thus, missense variants related to these pathways may disrupt biological processes.

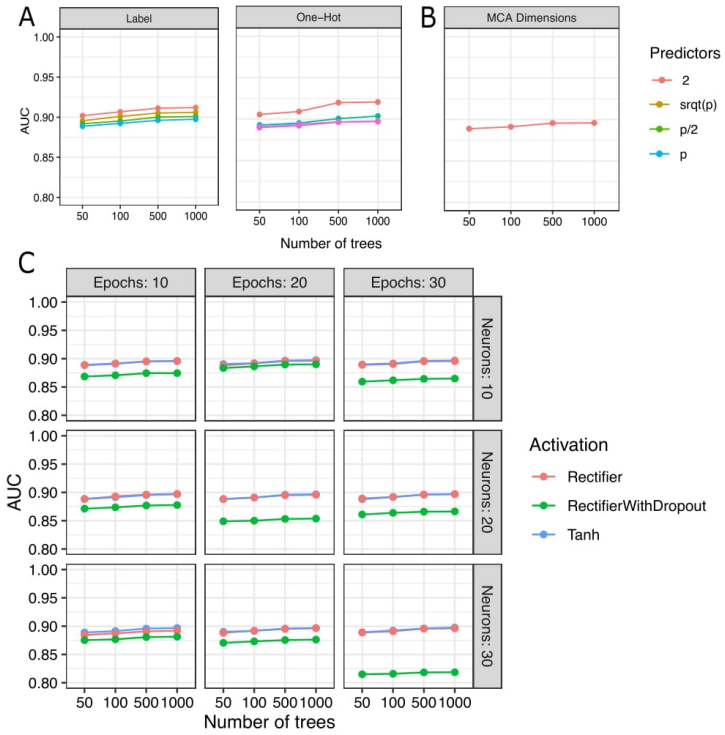



**Figure 2.** Fine-tuning analysis of Random Forest models. The Random Forest models were trained with label encoding and one hot encoding; learned data from multiple correspondence analysis and neural networks as autoencoders. (**A**) Random Forest shows high values of AUC when data is one hot encoded; (**B**) AUC results for Random Forest models trained with learned data from multiple correspondence analysis; (**C**) AUC results for Random forest models trained with autoencoded data.

### *3.4. Case Study: Alzheimer's Disease-Related Genes*

Genetic studies have identified candidate disease genes by mapping SNPs that may contribute to the development of dementia traits, such as Alzheimer's Disease (AD). Moreover, AD is a multifactorial and complex disease with a genetic basis that remains to be elucidated [6,33]. In ClinVar data, 18 SNPs (CI or VUS) are associated with AD. Prediction results show four pathogenic variants and 14 benign variants (see Figure 3B and Table 2). The A$\beta$ precursor protein (APP) gene shows 13 VUS and one CI variant. Two variants in the APP region were predicted as pathogenic, which may impact protein structure (NM_000484.4, c.982C>T, p.Arg328Trp) and (c.298C>T, p.Arg100Trp). MPO showed one pathogenic variant (c.1031G>A, p.Gly344Asp), as well PSEN1 (c.475TC, p.Tyr159His).

Molecular interactions between the aforementioned genes have been associated with AD. Extracellular formation of senile plaques, which are insoluble deposits of neurotoxic amyloid-$\beta$ (A$\beta$) peptides along with metal ions, is a histopathological hallmark of AD. Through redox reactions, metal ions are activated and may bond with A$\beta$ to catalyze Reactive Oxygen Species (ROS) such as hydroxyl, a highly reactive radical. This reaction may induce inflammation and oxidative damage to surrounding molecules [34–36].

Myeloperoxidase (MPO) is a myeloid enzyme abundant in neutrophil granulocytes and monocytes but not detectable in microglia. It plays a primary role in inflammatory and degenerative processes [37,38]. Studies reported the presence of MPO levels in the frontal cortex in A$\beta$ positive senile plaques and active microglia [38].

Mutations in presenilin-1 (PSEN1), presenilin-2 (PSEN2), and APP genes were previously described as a cause of autosomal-dominant early onset type AD [39,40] and familiar AD [41]. These genes are essential in the production of A$\beta$. APP encodes a precursor A$\beta$ protein, which is processed by the $\beta$-secretase and the $\gamma$-secretase complexes and leads to the production of A$\beta$. PSEN1 and PSEN2 encode presenilins, which constitute the catalytic subunit of the $\gamma$-secretase complex [39]. PSEN1 is also reported to cleave another type I transmembrane substrate, which could negatively affect notch signaling [41].

Opposite to APP, PSEN1, and PSEN2 mechanisms, A Disintegrin And Metalloprotease 10 (ADAM10) reduces the formation of A$\beta$ in physiological conditions and is associated with non-amyloidogenic and neuroprotective pathways [42]. ADAM10 encodes $\alpha$-secretase, a protein complex which cleaves the A$\beta$ region of APP, releasing a soluble fragment (sAPP$\alpha$) [43]. Previous studies have reported neuroprotective properties of sAPP$\alpha$ and proposed its enhancement as a therapeutic strategy for AD and other neurodegenerative diseases [44].

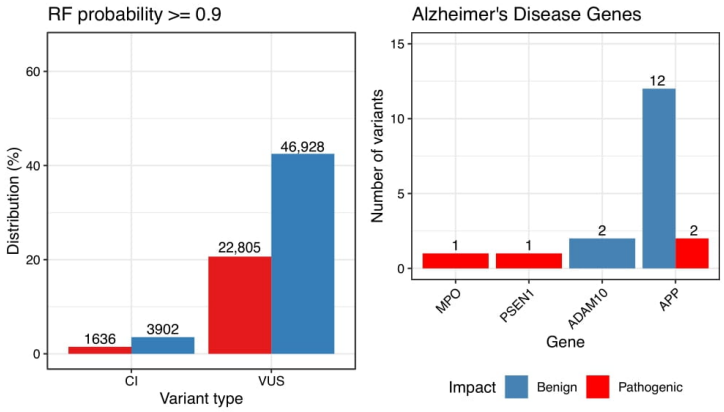



**Figure 3.** On the left, distribution of CI and VUS classified into benign and pathogenic after impact prediction with probability $\geq$ 0.9 by AmazonForest. On the right, distribution of variants for Alzheimer's Disease-related genes.

**Table 2.** AmazonForest prediction results for reclassification of genetic variants in genes associated with Alzheimer's disease.

| Chromosome | Position | Gene | Protein | Protein Change | dbSNP ID | ClinVar Significance | AmazonForest Prediction |
|---|---|---|---|---|---|---|---|
| 21 | 26000066 | APP | NM_000484.4 | c.982CT (p.Arg328Trp) | | VUS | Pathogenic |
| 21 | 26090000 | APP | NM_000484.4 | c.298CT (p.Arg100Trp) | rs200347552 | VUS | Pathogenic |
| 17 | 58278000 | MPO | NM_000250.2 | c.1031GA (p.Gly344Asp) | | VUS | Pathogenic |
| 14 | 73173702 | PSEN1 | NM_000021.4 | c.475TC (p.Tyr159His) | | VUS | Pathogenic |
| **Chromosome** | **Position** | **Gene** | **Protein** | **Protein Change** | **dbSNP ID** | **ClinVar Significance** | **AmazonForest Prediction** |
| 15 | 58665141 | ADAM10 | NM_001110.4 | c.541AG (p.Arg181Gly) | rs145518263 | VUS | Benign |
| 15 | 58665172 | ADAM10 | NM_001110.4 | c.510GC (p.Gln170His) | rs61751103 | VUS | Benign |
| 21 | 25997360 | APP | NM_000484.4 | c.1090CT (p.Leu364Phe) | rs749453173 | VUS | Benign |
| 21 | 25997413 | APP | NM_000484.4 | c.1037CA (p.Ser346Tyr) | | VUS | Benign |
| 21 | 26000018 | APP | NM_000484.4 | c.1030GA (p.Ala344Thr) | rs201045185 | VUS | Benign |
| 21 | 26000167 | APP | NM_000484.4 | c.881AG (p.Gln294Arg) | | VUS | Benign |
| 21 | 26021902 | APP | NM_000484.4 | c.803GA (p.Arg268Lys) | rs1601237753 | VUS | Benign |
| 21 | 26021954 | APP | NM_000484.4 | c.751GA (p.Gly251Ser | | VUS | Benign |
| 21 | 26021978 | APP | NM_000484.4 | c.727GA (p.Asp243Asn) | | VUS | Benign |
| 21 | 26022001 | APP | NM_000484.4 | c.704CT (p.Ala235Val) | | CI | Benign |
| 21 | 26022031 | APP | NM_000484.4 | c.674TC (p.Val225Ala) | rs746313873 | VUS | Benign |
| 21 | 26051060 | APP | NM_000484.4 | c.602CT (p.Ala201Val) | rs149995579 | VUS | Benign |
| 21 | 26051088 | APP | NM_000484.4 | c.574GA (p.Glu192Lys) | | VUS | Benign |
| 21 | 26170574 | APP | NM_000484.4 | c.47GA (p.Arg16Gln) | | VUS | Benign |

## 4. Discussion

In this study, we evaluated the performance of RF trained with encoding data and representation learning extracted from MCA and neural network-based autoencoders, aiming to produce a metaprediction model (AmazonForest). The best RF model with one hot encoding was chosen for (re)classification of VUS and CI variants. This study is the first to investigate different encoding methods and influences on pathogenicity predictions by RF and representation learning algorithms. We found that encoding methods and autoencoders had little influence on RF models (see ROC and AUC in Figure 2).

Metaprediction approaches were proposed based on distinct machine learning or statistical methods and differ in training datasets [9–11,13]. In fact, most of the reviewed metapredictors adopted decision tree-based methods, [9,11,13], which deal with categorical predictors without the need to reconstruct them [45]. However, all metapredictors are unclear about how they handle missing data, which may produce biased models. To avoid missing data bias and to obtain a reliable and robust model, our study removed variants with missing data from the training set. Thus, VUS and CI variants were reclassified if they showed data for the aforementioned eight predictors.

Our proposed model was used for the pathogenicity prediction of VUS and CI variants. After prediction, we identified a valuable set of 24,428 variants, at a RF probability >= 0.9, identifying a variant dataset with a high probability of being pathogenic. This information could further improve our understanding of well-known diseases, as well as clarify molecular mechanisms involved in rare disorders. Therefore, AmazonForest can help to obtain more careful and accurate analyses of variants of uncertain significance and CI. Finally, we provided an online tool and well-annotated R scripts for a better user experience of pathogenicity prediction of genetic variants as well as (re)classification of CI and VUS variants.

The proposed model was compared to other prediction algorithms such as SVM and NB [46–48]. These additional comparison experiments are found in the Supplementary Materials.

## 5. Conclusions

Our benchmark shows that AmazonForest, a Random Forest-based model, presents satisfactory prediction results (AUC $\geq$ 0.93) regarding categorical data and one hot encoded data from eight functional impact predictors. Furthermore, we provide a new reclassified database and a model for programmatic prediction of large genetic variant sets of VUS and CI variants. Geneticists may consider the AmazonForest genetic variant data, and the web tool, for annotation of genome-wide studies, disease model tests, and investigations of variants pathogenicity and their associations to complex diseases, as demonstrated for Alzheimer's disease.

## 6. Software Availability

AmazonForest is available online at: https://www2.lghm.ufpa.br/amazonforest. AmazonForest is constructed based on open source tools and all code is available at https://github.com/hpalheta/amazonforest. To use the metaprediction model we make availabe a R script, which are available on https://github.com/hpalheta/amazonforest/tree/master/meta_prediction/amazonforest.R. All data was accessed on 8 December 2021.

**Supplementary Materials:** The following are available online at https://www.mdpi.com/article/10.3390/biology11040538/s1, Figure S1: Gini impurity index for eight functional impact predictors; Figure S2: ROC curves for Naive Bayes, Random Forest and Support Vector Machine; Table S1: Reactome pathway enrichment analysis of genes mapped for VUS and CI genetic variants with pathogenicity probability equals to 0.9; Table S2: Accuracy, F1-score and mean AUC for Naive Bayes Random Forest and SVM.

**Author Contributions:** H.G.A.P. and G.S.d.A. handled data, performed computational experiments, and implemented the AmazonForest online platform. Â.R.-d.-S. and G.S.d.A. are senior authors that designed and supervised the study. H.G.A.P., W.G.G., L.M.B., A.R.d.S., M.d.R.M. and G.S.d.A. wrote and collaboratively reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Rede de Pesquisa em Genômica Populacional Humana (Biocomputacional—Protocol No. 3381/2013/CAPES/Brazil); Conselho Nacional do Desenvolvimento Científico e Tecnológico—CNPq Brazil (Â.R.S. was supported by CNPq/Productivity: 304413/2015-1), Fundação Amazônia Paraense de Amparo à Pesquisa—FAPESPA (No. BJT—2021/658671), Hydro (Project 4227 Hydro/UFPA/FADESP/Brazil), Coordenação de Aperfeioamento de Pessoal de Nível Superior (CAPES/Brazil) and Pró-Reitoria de Pesquisa (PROPESP/Brazil) of Universidade Federal do Pará (UFPA/Brazil). The funders had no role in the design of the study, collection, analysis, interpretation of the data or writing of the manuscript.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The training/test dataset was made available at https://github.com/hpalheta/amazonforest/blob/master/meta_prediction/clinvar.train.csv. The reclassified variant dataset was made available at https://github.com/hpalheta/amazonforest/blob/master/meta_prediction/clinvar.civus_new_pred.csv, which can be loaded easily in R environment. All data and models was accessed on 8 December 2021.

**Acknowledgments:** The authors would like do thank the funding agencies and public databases.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| NGS | Next Generation Sequencing |
| GWAS | Genome Wide Association Studies |
| SNP | Single Nucleotide Polymorphism |
| ML | Machine Learning |
| VUS | Variants of uncertain significance |
| CI | Conflit of interpretation |
| RF | Random Forest |
| ROC | Receiver Operating Curve |
| AUC | Area Under Curve |

## References

1. MacArthur, J.; Bowler, E.; Cerezo, M.; Gil, L.; Hall, P.; Hastings, E.; Junkins, H.; McMahon, A.; Milano, A.; Morales, J.; et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **2017**, *45*, D896–D901. [CrossRef] [PubMed]
2. Araújo, G.S.; Lima, L.H.C.; Schneider, S.; Leal, T.P.; da Silva, A.P.C.; Vaz de Melo, P.O.; Tarazona-Santos, E.; Scliar, M.O.; Rodrigues, M.R. Integrating, summarizing and visualizing GWAS-hits and human diversity with DANCE (Disease-ANCEstry networks). *Bioinformatics* **2016**, *32*, 1247–1249. [CrossRef] [PubMed]
3. Deng, N.; Zhou, H.; Fan, H.; Yuan, Y. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget* **2017**, *8*, 110635. [CrossRef] [PubMed]
4. Unoki, H.; Takahashi, A.; Kawaguchi, T.; Hara, K.; Horikoshi, M.; Andersen, G.; Ng, D.P.; Holmkvist, J.; Borch-Johnsen, K.; Jørgensen, T.; et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.* **2008**, *40*, 1098–1102. [CrossRef]
5. Araújo, G.S.; Souza, M.R.; Oliveira, J.R.M.; Costa, I.G. Random Forest and Gene Networks for Association of SNPs to Alzheimer's Disease. In *Brazilian Symposium on Bioinformatics*; Springer: Cham, Switzerland, 2013; pp. 104–115.
6. Souza, M.; Araujo, G.; Costa, I.; Oliveira, J.; Initiative, A.D.N. Combined genome-wide CSF A$\beta$-42's associations and simple network properties highlight new risk factors for Alzheimer's disease. *J. Mol. Neurosci.* **2016**, *58*, 120–128. [CrossRef]
7. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **2013**, *42*, D980–D985. [CrossRef]
8. Alzu'bi, A.A.; Zhou, L.; Watzlaf, V.J. Genetic variations and precision medicine. *Perspect. Health Inf. Manag.* **2019**, *16*, 1a.

9. Ranganathan Ganakammal, S.; Alexov, E. An Ensemble Approach to Predict the Pathogenicity of Synonymous Variants. *Genes* **2020**, *11*, 1102. [CrossRef]
10. Hassan, M.S.; Shaalan, A.; Dessouky, M.; Abdelnaiem, A.E.; ElHefnawi, M. Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity. *Genomics* **2019**, *111*, 869–882. [CrossRef]
11. Jaravine, V.; Balmford, J.; Metzger, P.; Boerries, M.; Binder, H.; Boeker, M. Annotation of Human Exome Gene Variants with Consensus Pathogenicity. *Genes* **2020**, *11*, 1076. [CrossRef]
12. Dong, C.; Wei, P.; Jian, X.; Gibbs, R.; Boerwinkle, E.; Wang, K.; Liu, X. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **2015**, *24*, 2125–2137. [CrossRef] [PubMed]
13. Do Nascimento, P.M.; Medeiros, I.G.; Falcão, R.M.; Stransky, B.; de Souza, J.E.S. A decision tree to improve identification of pathogenic mutations in clinical practice. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 50. [CrossRef] [PubMed]
14. Richards, S.; Aziz, N.; Bale, S.; Bick, D.; Das, S.; Gastier-Foster, J.; Grody, W.; Hegde, M.; Lyon, E.; Spector, E.; et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **2015**, *17*, 405–424. [CrossRef]
15. Shihab, H.A.; Gough, J.; Cooper, D.N.; Stenson, P.D.; Barker, G.L.; Edwards, K.J.; Day, I.N.; Gaunt, T.R. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **2013**, *34*, 57–65. [CrossRef] [PubMed]
16. Kumar, P.; Henikoff, S.; Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **2009**, *4*, 1073. [CrossRef] [PubMed]
17. Adzhubei, I.; Jordan, D.M.; Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **2013**, *76*, 7–20. [CrossRef]
18. Choi, Y.; Chan, A.P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **2015**, *31*, 2745–2747. [CrossRef]
19. Reva, B.; Antipin, Y.; Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **2007**, *8*, R232. [CrossRef]
20. Schwarz, J.M.; Cooper, D.N.; Schuelke, M.; Seelow, D. MutationTaster2: Mutation prediction for the deep-sequencing age. *Nat. Methods* **2014**, *11*, 361–362. [CrossRef]
21. Chun, S.; Fay, J.C. Identification of deleterious mutations within three human genomes. *Genome Res.* **2009**, *19*, 1553–1561. [CrossRef]
22. Abdi, H.; Williams, L.J. Principal component analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics 2.4*; Wiley: Hoboken, NJ, USA, 2010; pp. 433–459.
23. Team, H. Deep Learning, Neural Networks and Autoenconders. 2022. Available online: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html?highlight=autoencoder (accessed on 2 December 2021).
24. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv:1207.0580.
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
26. Strobl, C.; Boulesteix, A.L.; Zeileis, A.; Hothorn, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
27. Goldstein, B.A.; Hubbard, A.E.; Cutler, A.; Barcellos, L.F. An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genet.* **2010**, *11*, 49.
28. Ojala, M.; Garriga, G.C. Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **2010**, *11*, 1833–1863.
29. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
30. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
31. Wes McKinney. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 10–16 July 2010; pp. 56–61.
32. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
33. Brito, L.M.; Ribeiro-dos Santos, Â.; Vidal, A.F.; de Araújo, G.S. Differential expression and mirna–gene interactions in early and late mild cognitive impairment. *Biology* **2020**, *9*, 251. [CrossRef]
34. Cheignon, C.; Tomas, M.; Bonnefont-Rousselot, D.; Faller, P.; Hureau, C.; Collin, F. Oxidative stress and the amyloid beta peptide in Alzheimer's disease. *Redox Biol.* **2018**, *14*, 450–464. [CrossRef]
35. Alasmari, F.; Alshammari, M.A.; Alasmari, A.F.; Alanazi, W.A.; Alhazzani, K. Neuroinflammatory cytokines induce amyloid beta neurotoxicity through modulating amyloid precursor protein levels/metabolism. *BioMed Res. Int.* **2018**, 1–8. [CrossRef]
36. Miller, L.M.; Wang, Q.; Telivala, T.P.; Smith, R.J.; Lanzirotti, A.; Miklossy, J. Synchrotron-based infrared and X-ray imaging shows focalized accumulation of Cu and Zn co-localized with $\beta$-amyloid deposits in Alzheimer's disease. *J. Struct. Biol.* **2006**, *155*, 30–37. [CrossRef] [PubMed]
37. Ji, W.; Zhang, Y. The association of MPO gene promoter polymorphisms with Alzheimer's disease risk in Chinese Han population. *Oncotarget* **2017**, *8*, 107870. [CrossRef]

38. Reynolds, W.F.; Rhees, J.; Maciejewski, D.; Paladino, T.; Sieburg, H.; Maki, R.A.; Masliah, E. Myeloperoxidase polymorphism is associated with gender specific risk for Alzheimer's disease. *Exp. Neurol.* **1999**, *155*, 31–41. [CrossRef] [PubMed]
39. Lanoiselée, H.M.; Nicolas, G.; Wallon, D.; Rovelet-Lecrux, A.; Lacour, M.; Rousseau, S.; Richard, A.C.; Pasquier, F.; Rollin-Sillaire, A.; Martinaud, O.; et al. APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. *PLoS Med.* **2017**, *14*, e1002270. [CrossRef] [PubMed]
40. Oksanen, M.; Petersen, A.J.; Naumenko, N.; Puttonen, K.; Lehtonen, Š.; Olivé, M.G.; Shakirzyanova, A.; Leskelä, S.; Sarajärvi, T.; Viitanen, M.; et al. PSEN1 mutant iPSC-derived model reveals severe astrocyte pathology in Alzheimer's disease. *Stem Cell Rep.* **2017**, *9*, 1885–1897. [CrossRef] [PubMed]
41. Arber, C.; Lovejoy, C.; Harris, L.; Willumsen, N.; Alatza, A.; Casey, J.M.; Lines, G.; Kerins, C.; Mueller, A.K.; Zetterberg, H.; et al. Familial Alzheimer's disease mutations in PSEN1 lead to premature human stem cell neurogenesis. *Cell Rep.* **2021**, *34*, 108615. [CrossRef] [PubMed]
42. Yuan, X.Z.; Sun, S.; Tan, C.C.; Yu, J.T.; Tan, L. The role of ADAM10 in Alzheimer's disease. *J. Alzheimer's Dis.* **2017**, *58*, 303–322. [CrossRef]
43. Manzine, P.R.; Ettcheto, M.; Cano, A.; Busquets, O.; Marcello, E.; Pelucchi, S.; Di Luca, M.; Endres, K.; Olloquequi, J.; Camins, A.; et al. ADAM10 in Alzheimer's disease: Pharmacological modulation by natural compounds and its role as a peripheral marker. *Biomed. Pharmacother.* **2019**, *113*, 108661. [CrossRef]
44. Spilman, P.; Bredesen, D.; John, V. Enhancement of sAPPalpha as a Therapeutic Strategy for Alzheimer's and other Neurodegenerative Diseases. *J. Alzheimer's Neurodegener. Dis.* **2015**, *1*, 1–10.
45. Au, T.C. Random forests, decision trees, and categorical predictors: The "absent levels" problem. *J. Mach. Learn. Res.* **2018**, *19*, 1737–1766.
46. Mitchell, T.M. Generative and discriminative classifiers: Naive bayes and logistic regression. *Mach. Learn.* **2010**, 1–17.
47. IJCAI 2001 workshop on empirical methods in artificial intelligence. *Mach. Learn.* **2001**, *3*, 41–46.
48. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

biology

MDPI

*Article*

# Supplementary Material - AmazonForest: In Silico Metaprediction of pathogenic variants

**1. Feature Importance**

Feature importance were accessed for the best Random Forest model combined with label encoder. The Figure S1 shows the distribution of Gini impurity index for the eight predictors.

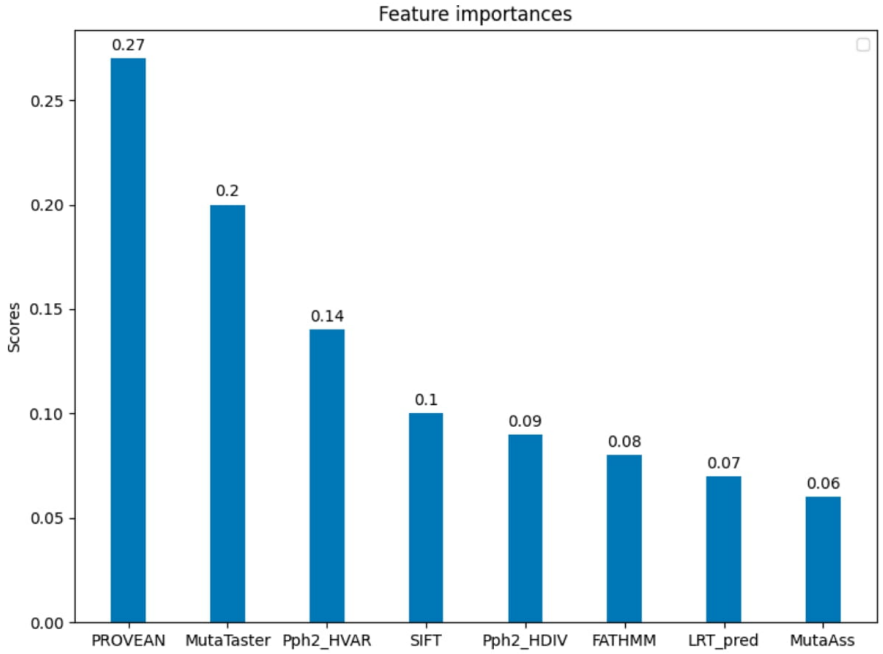



**Figure S1.** Gini impurity index for eight functional impact predictors.

**2. Enriched Pathways**

| Biological Pathway | Genes | False–Discover Rate |
|---|---|---|
| Diseases of metabolism | 76 | 3.19E-07 |
| Extracellular matrix organization | 56 | 1.40E-04 |
| Muscle contraction | 42 | 0.00200118 |
| Degradation of the extracellular matrix | 37 | 4.41E-06 |
| Diseases of glycosylation | 35 | 0.0048245 |
| Cardiac conduction | 30 | 0.00821261 |
| Signaling by PDGF | 26 | 1.15E-05 |
| ECM proteoglycans | 25 | 1.07E-05 |
| Integrin cell surface interactions | 25 | 4.56E-05 |

| | | |
|---|---|---|
| Collagen formation | 25 | 4.92E-04 |
| Collagen degradation | 24 | 4.41E-06 |
| Signaling by MET | 23 | 3.34E-04 |
| Collagen biosynthesis and modifying enzymes | 22 | 1.40E-04 |
| Assembly of collagen fibrils and other multimeric structures | 21 | 9.67E-05 |
| NCAM signaling for neurite out-growth | 21 | 1.40E-04 |
| Collagen chain trimerization | 20 | 1.97E-06 |
| Non-integrin membrane-ECM interactions | 19 | 2.27E-04 |
| NCAM1 interactions | 16 | 2.27E-04 |
| MET promotes cell motility | 14 | 0.0035418 |
| MET activates PTK2 signaling | 13 | 5.12E-04 |
| Laminin interactions | 12 | 0.00177183 |
| Signaling by PDGFRA extracellular domain mutants | 9 | 0.00355751 |
| Signaling by PDGFRA transmembrane, juxtamembrane and kinase domain mutants | 9 | 0.00355751 |
| Anchoring fibril formation | 8 | 0.00399627 |

**Table S1.** Reactome pathway enrichment analysis of genes mapped for VUS and CI genetic variantswith pathogenicity probability equals to 0.9.

**3. Performance of Naive Bayes, Random Forest and Support Vector Machines**

The model evaluation of Random Forest models indicates stability concerning AUC values, close to 0.93, as shown in the main text (see, Section 3.2, Figure 2). In this way, we chose the RF model trained with one-hot to perform further analyses. For comparison purposes, we performed experiments with the Naïve Bayes method and Support Vector Machines.

Naive Bayes (NB) is a probabilistic classifier algorithm, which assumes that all features are independent, so one feature does not influence other [46]. NB is based on the Bayes Theorem. Thus a simplistic classifier that has been reported to have an excellent performance regarding execution time and accuracy, even when considering analyses of large datasets [47]. The NB finds the probability of a given set of predictors for all possible values of the class variable $y$ and finds the maximum probability. This procedure can be expressed mathematically as: $y = argmax_y P(y) \prod^n P(x_i|y)$. In the experiment with NB, we kept the default parameters available in the scikit-learn library (https://scikit-learn.org/stable/modules/naive_bayes.html, accessed on 6 February 2022).

Support Vector Machine (SVM) is a machine learning method based on statistical learning theory.Proposed by Vapnik [48], SVM has been applied for classification and regression problems. SVM performs the classification of samples based on a set of hyperplanes estimated from high-dimensional datasets. If samples are not linearly separable, a nonlinear transformation is performed. SVM takes advantage of kernel functions for this purpose. Kernels such as the radial basis function, linear, polynomial, and sigmoid were used in the training step to select support vectors that outline the best hyperplane in the feature space. The implementation is available at https://scikit-learn.org/stable/modules/svm.html,accessed on 6 February 2022.

In general, the evaluation results of Random Forest, Naive Bayes, and SVM models indicate good performance (AUC > 0.9), as can be seen in Table 2 and Figure 2. However, the SVM model trained linear kernel yields similar results to Random Forest. The AUC for the two models corresponds to 0.93 with a standard deviation of 0.01. Thus, we chose Random Forest to compose the AmazonForestbase, in contrast to the SVM, which

deals better with non-categorical data, has a costly computational complexity and training time for large databases.

**Table S2.** Accuracy, F1-score and mean AUC for Naive Bayes Random Forest and SVM.

| Model | Accuracy | F1-Score | Mean AUC |
|---|---|---|---|
| Naive Bayes | 0.844 | 0.840 | 0.92 |
| Random Forest | 0.861 | 0.852 | 0.93 |
| SVM- RBF | 0.858 | 0.850 | 0.90 |
| SVM - Linear | 0.861 | 0.854 | 0.93 |
| SVM- Poly | 0.856 | 0.849 | 0.91 |
| SVM - Sigmoid | 0.813 | 0.795 | 0.88 |

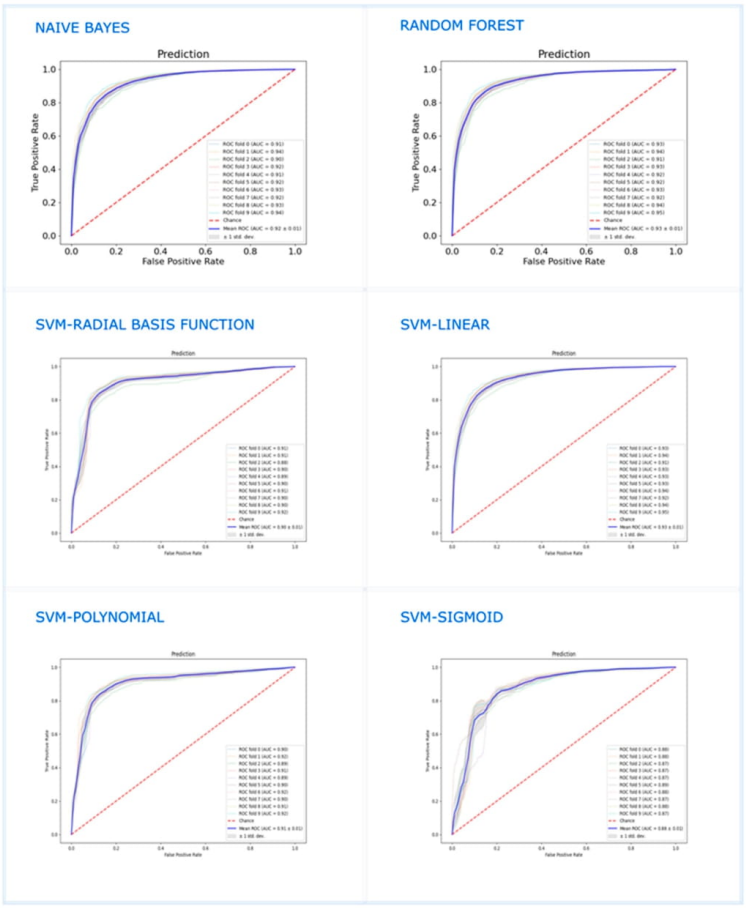



**Figure S2.** ROC curves for Naive Bayes, Random Forest and Support Vector Machine.

## References

46. Mitchell, T.M. Generative and discriminative classifiers: Naive bayes and logistic regression. *Machine learning* **2010**, pp. 1–17.
47. Rish, I.; others. An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, Vol. 3, pp. 41–46.
48. Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.

# 4 CAPÍTULO II- PATOGENICIDADE *IN SILICO* PARA ESCALA GENÔMICA: APLICAÇÕES NO CÂNCER

## 4.1 INTRODUÇÃO

A construção do conhecimento básico em pesquisas que envolvem o entendimento dos mecanismos e/ou processos de adoecimento possui alto grau de importância, pois é a primeira etapa para a transformação desse conhecimento em pesquisa translacional. Portanto, a identificação de métodos preventivos de diagnósticos ou de tratamento de doenças complexas, como o câncer, ainda são tarefas difíceis e pouco estabelecidas.

Os estudos das diferentes plataformas de sequenciamento, como do genoma completo (MARDIS, 2013), exoma (SHYR; KUSHNIRUK; WASSERMAN, 2014), e da associação ampla do genoma (GWAS), sugerem que as diferentes manifestações do câncer têm uma arquitetura gênica complexa e muitas vezes multifatorial. Portanto, o acúmulo de dados biológicos e o conhecimento construído sobre os elementos moleculares envolvidos no processo de cancerização, perpassam necessariamente pelos distintos estudos publicados em bancos de dados, a maioria de domínio público, e pela análise de complexidade das diferentes abordagens de bioinformáticas (GAO *et al.*, 2019; CAMPBELL *et al.*, 2018).

### 4.1.1 DIVERSIDADE HUMANA

As variantes genéticas humanas não estão distribuídas uniformemente no mundo, mas agrupadas em grupos populacionais em função da história compartilhada e separação geográfica das populações. Nesse sentindo, compreender não somente as consequências funcionais, como também a distribuição geográfica dessas variantes, contribui para o melhor entendimento da história das populações humanas e o desenvolvimento de melhores estratégias para prática e aplicabilidade na área da saúde (SANTOS *et al.*, 2013).

A conclusão do sequenciamento do primeiro genoma humano (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM, 2001; VENTER et al., 2001; NURK et al., 2022), e o avanço das tecnologias de sequenciamento promoveram diversos projetos que buscaram expandir o conhecimento das variantes genéticas humanas e sua distribuição nas populações humanas. Entre esses projetos destacam-se: HapMap (FRAZER et al., 2007); *1,000 Genomes* (THE 1000 GENOMES PROJECT CONSORTIUM et al., 2015); ExAC e gnomAD (LEK et al., 2016); e *Simons Genome Diversity Project* (MALLICK et al., 2016), os quais sequenciaram e disponibilizaram os dados de genoma e exoma para coortes de populações humanas ao redor do mundo.

Dessa forma, após o fim da era genômica e da finalização do sequenciamento referência do genoma humano, em março de 2022 (NURK *et al.*, 2022), novas metas e paradigmas surgiram, dentre os quais citamos - na genética humana: (1) o desenvolvimento de novas ferramentas de bioinformática; (2) a busca intensa pelo entendimento dos mecanismos regulatórios envolvidos nas doenças; (3) a melhoria dos métodos de diagnósticos e de prognósticos moleculares mais precisos e específicos nas diversas doenças que afligem a humanidade; e (4) identificação de possíveis alvos para o desenvolvimento de novas drogas aplicadas nos tratamentos terapêuticos.

Com intuito de entender os mecanismos genéticos de diferentes doenças e características estão sendo desenvolvidos diversos estudos de associação genômica (GWAS), dos quais 27.500 estudos de 550 publicações e mais de 100k associações de SNPs estão catalogadas no *GWAS Catalog*, disponível em https://www.ebi.ac.uk/gwas/ (HAYHURST *et al.*, 2022). Entretanto, a maioria dos estudos desenvolvidos foi realizado em coortes de ancestralidade predominantemente europeia. O que resulta na baixa diversidade, representatividade das populações africanas, asiáticas e povos nativos americanos.

Em 2016, apenas 19% dos participantes investigados em plataformas NGS possuía ancestria não-europeia, embora não se tenha conhecimento de que indivíduos de ancestria africana e hispânica possuem um número desproporcional de associações (POPEJOY & FULLERTON, 2016). Além disso, a aplicação desses resultados em populações diferentes das estudadas é limitado e resultam em resultados enviesados (MARTIN et al., 2017; KIM et al., 2018). Esse fato destaca a importância da inclusão desses grupos populacionais, assim como o estudo de populações miscigenadas como a brasileira. A aplicabilidade limitada dos resultados de *GWAS* em populações não-europeias é resultado das distintas histórias evolutivas e genéticas de cada população humana. Projetos como *1,000 Genomes* demonstraram que a maioria das mutações são raras (MAF < 1%) e específicas do grupo (*THE 1,000 GENOMES PROJECT CONSORTIUM* et al., 2015). Estudos focados em um único grupo ancestral podem levar a perda de mutações funcionais, como por exemplo, a mutação de perda de função (LoF) no gene *PCSK9* que resulta em baixos níveis de colesterol em descendentes de africanos (COHEN et al., 2005). Esse foco também pode resultar em erros de interpretação, como mutações descritas como patogênicas para cardiomiopatias hipertrófica e prevalente em populações africanas, tornando improváveis as causas da doença.

Com base nos dados de sequenciamento de 33 tipos distintos de câncer depositados na plataforma TCGA (<https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>), Gao et al. (2019) e Campell et al. (2018) realizaram análises integrativas e destacaram alterações genômicas de potencial interesse biológico ou terapêutico relacionados aos diversos tumores (GAO et al., 2019; CAMPBELL et al., 2018).

### 4.1.2 VARIABILIDADE GENÉTICA HUMANA

A variabilidade genética é produto do processo histórico de expansão, migração e isolamento das populações humanas, dessa maneira que as populações desenvolveram seus próprios perfis genéticos ao longo de sua formação, com perdas e ganhos de novos alelos.

Atualmente, estão catalogadas 660 milhões de variantes genéticas humanas no dbSNP (build 151), principal base de variantes genéticas, das quais apenas 381 milhões estão contidas nas regiões gênicas. Esse extenso catálogo de variantes genéticas tornou-se possível graças ao recente avanço da tecnologia de sequenciamento. Usando destas ferramentas, grandes projetos exploraram a variabilidade genética do genoma e exoma de diversas populações humanas saudáveis ao redor do mundo caracterizando as mutações e avaliando suas frequências alélica.

Dentre esses projetos destacam-se o *1,000 Genomes Project* (THE 1000 GENOMES PROJECT CONSORTIUM et al., 2015a), *Exome Aggragation Consortium* (ExAC), *Genome Agregation Database* (gnomAD) (LEK et al., 2016) e (iv) *Simons Genome Diversity Project* (MALLICK et al., 2016). Tais projetos disponibilizaram seus dados para o público e compreendem indivíduos de populações Europeias, Africanas, Asiáticas, Oceânicas e Americanas (Tabela1).

Tabela 1: Amostras investigadas por projetos genômicos por região geográfica.

| Região Geografica | 1,000 Genomes Project | Simons Genome Diversity Project | ExAC | gnomAD |
|---|---|---|---|---|
| Europeia | 503 | 71 | 36677 | 76266 |
| Africana | 661 | 39 | 5203 | 12020 |
| Sul Asiática | 489 | 38 | 8256 | 15391 |
| Sudeste Asiática | 504 | 49 | 4327 | 9435 |
| Americanos Miscigenados | 347 | 0 | 5789 | 17210 |
| Nativo Americanos | 0 | 23 | 0 | 0 |
| Outros | 0 | 54 | 454 | 8310 |
| Total | 2504 | 274 | 60706 | 138632 |

Embora extenso, os dados genéticos disponíveis e a interpretação dos efeitos biológicos das variantes ainda é complexa, eles variam de acordo com o contexto genético na qual estão presentes. Desse modo, em regiões codificantes (regiões dos genes transcritas em RNA mensageiro) podem ser classificadas em (Figura 5): (i*) silent* ou silenciosas, quando não afetam a sequência de aminoácidos inscrita no gene; (ii) *misense,* quanto trocam o aminoácido codificado; (iii) *nonsense,* quando ocasiona a transcrição de uma proteína truncada ou não funcional pela retira ou inclusão de um códon de parada; (iv) *frameshift*, quando altera a matriz de leitura dos códons pela inserção ou deleção de algumas bases; e (iv) *splicing*, quando afeta o processo de *splicing* ocasionando a perda de éxons ou retenção de introns.

As mutações *nonsense* e *frameshift* geralmente ocasionam a perda de função da proteína (LoF) e são, portanto, mais fáceis de interpretar. Enquanto as mutações misense possuem impacto incerto quanto a função do gene, sendo necessários estudos extensos para interpretá-las.

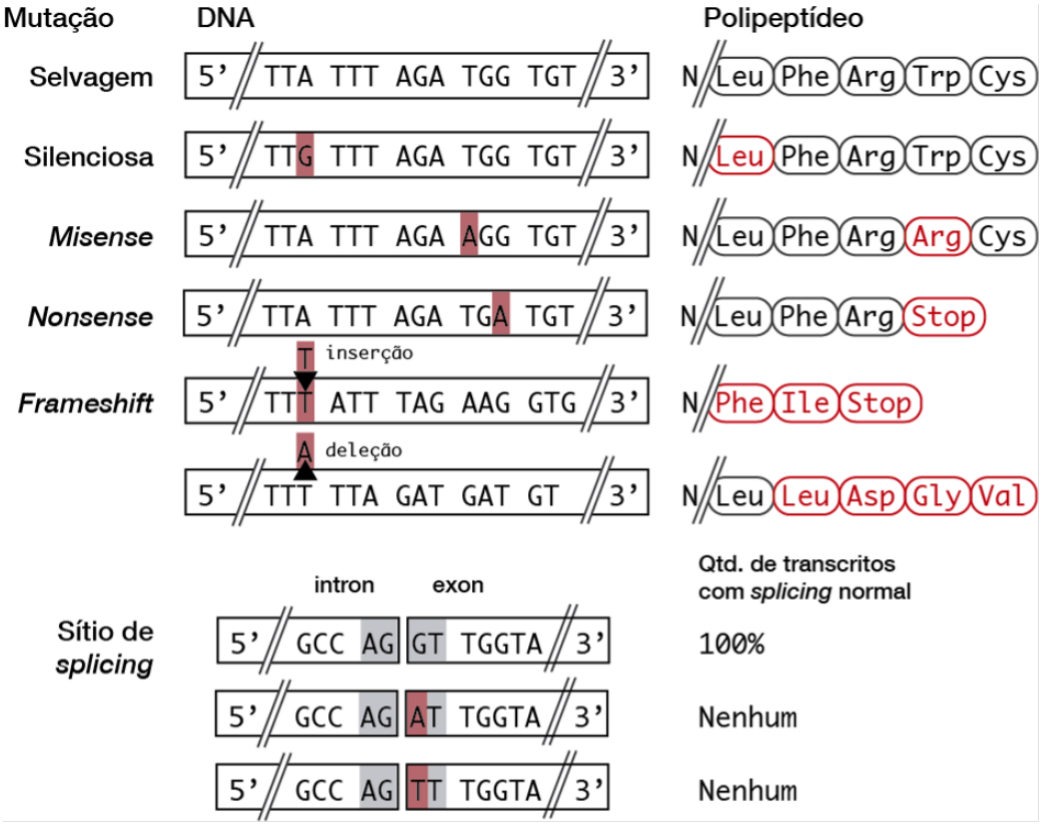



Figura 5: Mutações em regiões codificantes e seus efeitos na sequência de aminoácidos (SANTOS, 2014).

Por outro lado, determinar as consequências e o gene afetado em regiões não codificantes (e.g. introns, intergênica e promotoras) é muito mais complexo. Essas regiões compreendem 98% do genoma humano e acumulam o maior volume de variantes genéticas, as quais influenciam principalmente a expressão gênica, afetando a afinidade de ligação dos fatores de transcrição, fatores epigenéticos, acessibilidade e organização tridimensional do material genético sendo, portanto, componentes essenciais dos mecanismos genéticos de características e doenças complexas (FARNHAM, 2009; ZHANG & LUPSKI, 2015; KHURANA et al., 2016).

A interpretação dessas variantes dá-se principalmente por estudos funcionais específicos ou estudos de associação genômico (GWAS). Tais estudos buscam associar traços fenotípicos ou doenças à variantes ao longo do genoma, entretanto, a maioria dos resultados

identificam mutações em regiões não-codificantes, que são predominantemente localizados em sítios regulatórios.

Além do contexto genômico onde as variantes se encontram, outras ferramentas importantes para interpretação clínica, tais como estudos de associação e diagnóstico de diversas doenças, são as bases de dados de frequência alélica (como gerado pelos projetos genômicos) e catálogos de interpretação clínica (RICHARDS et al., 2015).

Um critério comum para interpretação clínica de variantes é a frequência alélica em populações saudáveis, mutações muito comuns ou estabelecidas que, geralmente, não são patogênicas. Adicionalmente, o banco de dados ClinVar reúne a interpretação clínica de aproximadamente 437 mil variantes (disponível em: https://www.ncbi.nlm.nih.gov/clinvar) sendo uma ferramenta essencial para diagnóstico clínico.

Por outro lado, a interpretação do impacto das variantes precisa ser contextualizada na população de origem dos pacientes estudados. Por exemplo, esquemas de riscos poligênicos superestima ou subestimar o risco de pacientes quando implementado em pacientes de origem diferente das estudadas (KIM et al., 2018).

As populações humanas são diversificadas tanto no que diz respeito a aspectos históricos e culturais, quanto genéticos. Tal fato é fruto de milhares de anos de expansão, migração e adaptação. As populações possuem histórias genéticas únicas e distintas arquiteturas genéticas com diferentes perfis de mutações, frequência alélicas e desequilíbrio de ligação. Tudo isso produz diferentes estratégias adaptativas e distintas respostas aos estímulos e riscos.

Considerando o exposto, o desenvolvimento de novas ferramentas de bioinformática torna-se imprescindível porque elas podem ser aplicadas na busca de soluções reais ou podem ajudar no entendimento dos mecanismos envolvidos nas doenças, como o câncer. No câncer especificamente, existem diversas variantes genéticas presentes em nosso genoma (do tipo

germinativa ou somática) que são responsáveis ou envolvidas no processo de cancerização. Portanto, o seu entendimento é de fundamental importância para estabelecer uma base biológica relacionada ao tipo e ao número de variantes presentes em um indivíduo. De igual forma, também há necessidade da interpretação da manifestação clínica que essas variantes podem produzir.

Diante do contexto apresentado, o presente trabalho teve como objetivo analisar o panorama de diversidade genética populacional de SNPs não-sinônimos, em vias específicas de câncer, bem como realizar uma análise de distribuição das frequências desses SNPs em populações de origem Africana, Europeia e Latino-Americana, relacionadas a câncer, a partir da mineração de bases de dados públicas e consolidadores funcionais de patogenicidade.

## 4.2 MATERIAL E MÉTODOS

O fluxo computacional/bioinformático desse trabalho consistiu das seguintes etapas: i) extração e pré-processamento de dados funcionais de patogenicidade para variantes em escala genômica (ver Seção 4.2.1); ii) consolidação da predição de patogenicidade por meta-preditor (ver Seção 4.2.2); iii) integração da meta-predição a vias biológicas relacionadas a câncer (Seção 4.2.3); e iv) investigação do panorama populacional das variantes genéticas (Seção 4.2.1) (Figura 6).

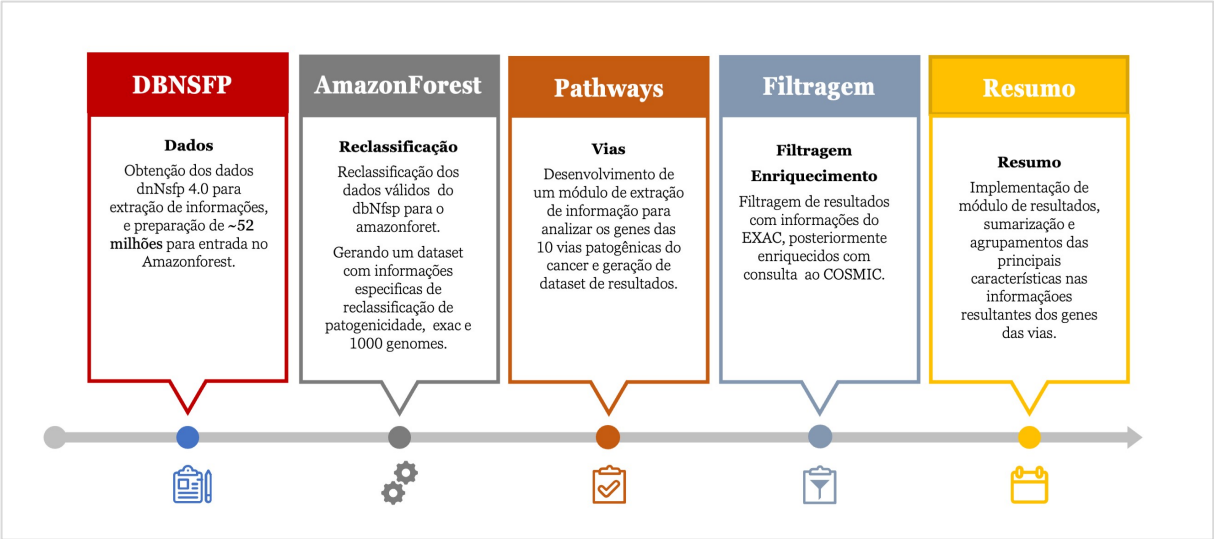



Figura 6: Fluxo computacional das etapas investigadas.
Fonte: Autor.

### 4.2.1 PREDIÇÃO FUNCIONAL DE VARIANTES GENÉTICAS NÃO-SINONIMAS

No trabalho desenvolvido as anotações funcionais de patogenicidade foram extraídas para cada gene. Para essa etapa, utilizamos o SnpEff(v.4.3) e o dbNSPF4.0, enquanto ferramentas, e um banco de dados desenvolvido para apoio a análises funcionais de variantes não-sinônimas (nsSNVs). O dbNSPF tem como base o Gencode/Ensembl e contem aproximadamente um total de 84 milhões de variantes registradas (consultado em <https://drive.google.com/file/d/1XQI2m_403yq-TLxJ1QHtkzKE7_c_9Gal/view> em <12/05/2020>). O dbNSFP mantém resultados de patogenicidade de 38 algoritmos preditores, dos quais extraímos informações do *SIFT, SIFT4G*, *Polyphen2-HDIV, Polyphen2-HVAR*, *LRT, MutationTaster2, MutationAssessor, FATHMM, PROVEAN*. Esses preditores foram escolhidos, pois são base para integração com o AmazonForest que consolida a predição de patogenicidade como um meta-preditor, descrito na próxima seção.

Além das informações sobre preditores de patogenicidade, o *dbNSFP* inclui frequências alélicas observadas do Projeto *1,000 Genomes, UK10K*, dados do consórcio *ExAC*, dados do banco de dados *gnomAD* e o Projeto de sequenciamento de exomas *NHLBI*

(LIU *et al.*, 2020). As informações de frequência alélica serão base para extensão das análises de genômica de populações.

### 4.2.2 META-PREDIÇÃO E CONSOLIDAÇÃO DE PATOGENICIDADE

Utilizamos o AmazonForest (PALHETA *et al.*, 2022) um modelo de meta-predição utilizado para consolidar as predições funcionais, com base na anotação de oito preditores categóricos de impacto funcional (*FATHMM, SIFT, PolyPhen-2 (HDIV), PolyPhen-2 (HVAR), PROVEAN, MutationAssessor, MutationTaster2* e *LRT*). O AmazonForest implementa o melhor modelo de metapredição baseado na estratégia de codificação *one-hot* com *Random Forest* (AUC = 0,93).

### 4.2.3 GENÉTICA DE VIAS BIOLÓGICAS CANÔNICAS EM CÂNCER

Neste trabalho investigamos mutações genéticas em dez vias canônicas relacionadas a câncer definidas em (SANCHEZ-VEGA *et al.*, 2018). Partindo de experimentos de RNA-Seq do TCGA), o pesquisador reportou dez vias comuns relacionadas a 33 tipos de tumores em 9.125 amostras e analisou os mecanismos e padrões de variantes somáticas em dez vias canônicas: Ciclo Celular, Hippo, Myc, Notch, Nrf2, PI3-quinase/Akt, RTK-RAS, sinalização de TGF$\beta$, p53 e $\beta$-cateninaWnt.

A Figura 7 representa cada via reportada e seus elementos moleculares. O fluxo das análises computacionais e a feitas por especialistas nas quais definiram as dez vias citadas seguiu as seguintes etapas:

i) Uma definição de escopo inicial para obtenção dos genes foi baseada em informações curadas em publicações previamente no TCGA, obtenção e tratamentos dos dados do TCGA PanCancer Atlas, consultas às base de dados de literaturas cientificas e revisão da literatura;

ii) Uma análise feita por um grupo de revisão ou por um especialista em estudo de determinada via, utilizando um conjunto de definições de alterações drivers e aplicações de ferramentas como *MutSigCV* (LAWRENCE *et al.*, 2014), *GISTIC 2.0* (MERMEL *et al.*, 2011), *OncoKB* (CHAKRAVARTY *et al.*, 2017) etc;

iii) finalizando com análises utilizando *cBioPortal* (CERAMI *et al.*, 2012), *SELECT* (MINA *et al.*, 2017) e *PathwayMapper* (BAHCECI *et al.*, 2017). Assim os autores pelo estudo chegaram nas dez 10 Vias (*pathways*) de sinalização oncogênicas TCGA que descrevemos resumidamente abaixo:

- Ciclo Celular - Regulação da progressão do ciclo celular mitótico envolvendo uma cascata de sinalização de ciclinas e quinases dependentes de ciclina, bem como vários pontos de verificação regulatórios. Tendo por genes representativos, *CDKN2A/B*, familia *CCND*, familia *CDK*.

- HIPPO - Envolvido no controle do tamanho do órgão. Central para esta via é a regulação dos co-ativadores de transcrição *YAP/TAZ* que promovem a transcrição de genes envolvidos na proliferação celular. Tendo por genes representativos, *LATS1/2*, *YAP1*, *TAZ* (*WWTR1*).

- *MYC* - Envolve uma série de complexos de regulação da transcrição: *MYC-MAX*, *MAX-MXD*, *MAX-MGA* e o sensor de energia, complexo MondoA-Mlx na regulação da resposta apoptótica e diferenciação celular. Sendo os genes representativos, *MYC*, *MAX*, *MGA*.

- *NOTCH* - Via envolvida na comunicação célula-célula, destino da célula. A clivagem de receptores Notch leva ao deslocamento de um complexo repressor de transcrição no *RBPJ* (um fator de transcrição também conhecido como CSL) acompanhado pelo recrutamento

de um complexo de ativação (incluindo *MAMLs*) que leva à transcrição de genes alvo *NOTCH*. Sendo os genes representativos, família *NOTCH*, família *JAG*, *EP300*, *CREBBP*.

- *NRF2* - Envolve a regulação do fator de transcrição *NFE2L2* por *KEAP1*. O *NFE2L2* regula os genes com os elementos de resposta antioxidante (ARE) que auxiliam na resposta celular contra o estresse oxidativo e ajudam na quimiorresistência ao câncer. Tendo por genes representativos, *NFE2L2*, *KEAP1*, *CUL3*.

- *PI3K* - Uma cascata de sinalização envolvendo a fosforilação de *PI3K* de *AKT* levando à ativação do complexo mTORC1. O mTORC1 funciona como um sensor metabólico e controla a abundância de proteínas, afetando os processos envolvidos na produção de proteínas e na tradução de RNA, levando a alterações no crescimento e sobrevivência celular. Tendo por genes representativos, *NFE2L2*, *KEAP1*, *CUL3*.

- *RTK/RAS* - Uma via de cascata de sinalização iniciada pela ativação de RTKs seguiu a transdução de sinal através de Ras, depois Raf e, em seguida, membros da família *MEK*. Essa cascata leva à ativação de vários fatores de transcrição que regulam processos envolvendo proliferação e sobrevivência celular. Tendo por genes representativos família *RTK*, *RAS*, *BRAF*, *MAP2K1*, *NF1*.

- *TGFβ* - Rede de sinalização envolvida no crescimento, proliferação, apoptose e diferenciação envolvendo a ativação de receptores de TGF*beta* pela citocina TGF*beta* que leva à ativação da transcrição gênica por SMADs. Tendo por genes representativos, família *SMAD*, *TGFBR1/2*, *ACVR2A/B*.

- *TP53* - Caminho centrado em torno da regulação do supressor tumoral *TP53*, um gene que regula a apoptose, parada do ciclo celular, senescência e reparo de DNA. Tendo por genes representativos *TP53*, *CDKN2A*, *ATM*, *MDM2/4*.

- *WNT* - Envolvido tanto no desenvolvimento quanto na homeostase tecidual. A via canônica Wnt envolve a transdução de sinal iniciada pela ligação do ligante Wnt aos receptores da família Frizzled, levando à desregulação da degradação da beta-catenina e, finalmente, à indução da transcrição via fatores de transcrição TCF/LEF pela betacatenina. Tendo por genes representativos, *APC*, *CTNNB1*, família *FZD*, *RNF43*.

Na Figura 7 é apresentado as 10 vias com seus respectivos detalhes e genes representativos de cada via.

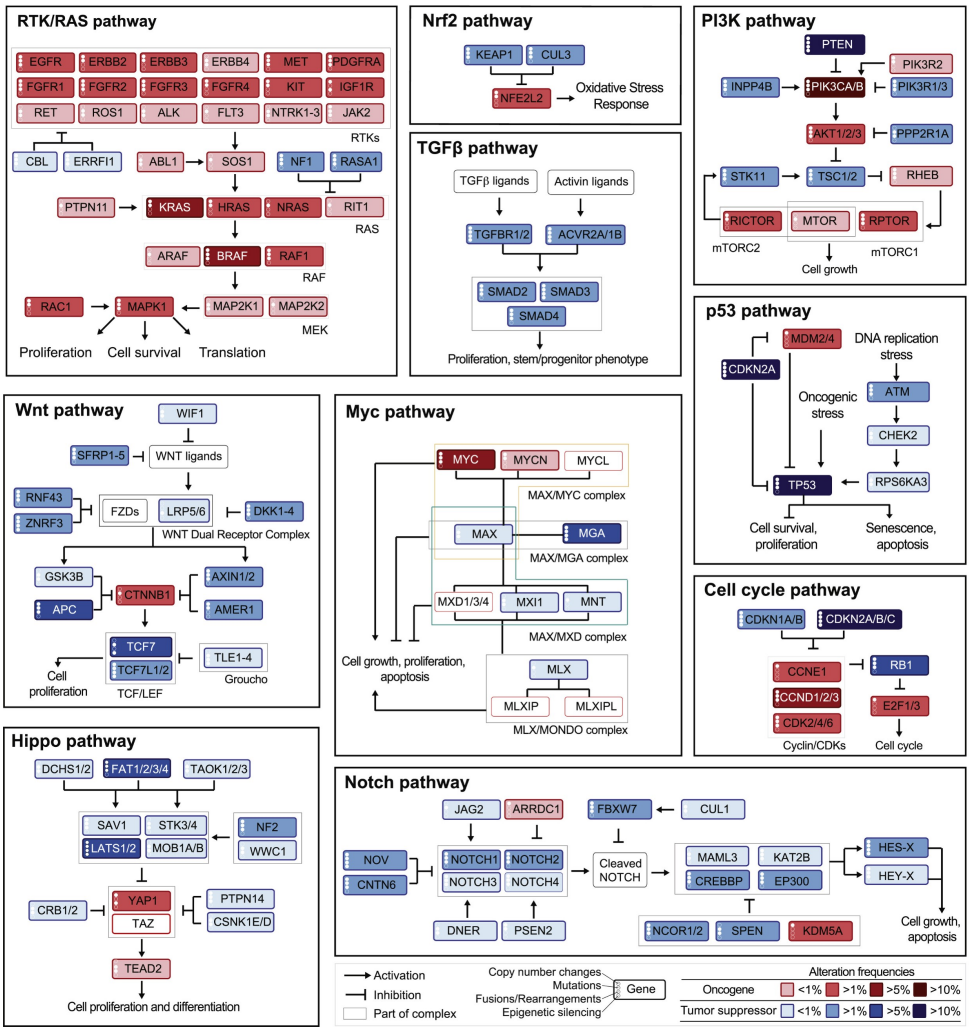



Figura 7: As 10 vias (pathways) de sinalização oncogênicas a partir de dados do TCGA com seus respectivos genes (SANCHEZ-VEGA et al., 2018).

## 4.3 RESULTADOS

De modo geral, investigamos a patogenicidade de 52 milhões variantes catalogadas no dbNSFP. Notamos que uma porcentagem considerável da base de dados *dbNSFP* possuem dados faltantes para preditores funcionais. As distribuições de dados faltantes e variantes anotadas podem ser visualizada por cromossomo na Figura 8A. Com esse diagnóstico, investigamos as variantes que apresentam predições dos oito preditores de impacto funcional, utilizados como entrada para classificação com o modelo AmazonForest. Aproximadamente 48,6 milhões (93,35%) de variantes apresentam-se como benignas e 3,4 milhões (6,65%) com alta suspeita de patogenicidade. As predições foram restritas com probabilidade de >= 0.95. A distribuição da meta-predição pelo AmazonForest por cromossomo pode ser conferida na Figura 8B. Ressaltamos que das variantes com alta suspeita de patogenicidade, 3,4 milhões estavam catalogadas no *dbNSFP* como variantes sem significado (VUS) e 1,113 com conflito de interpretação.

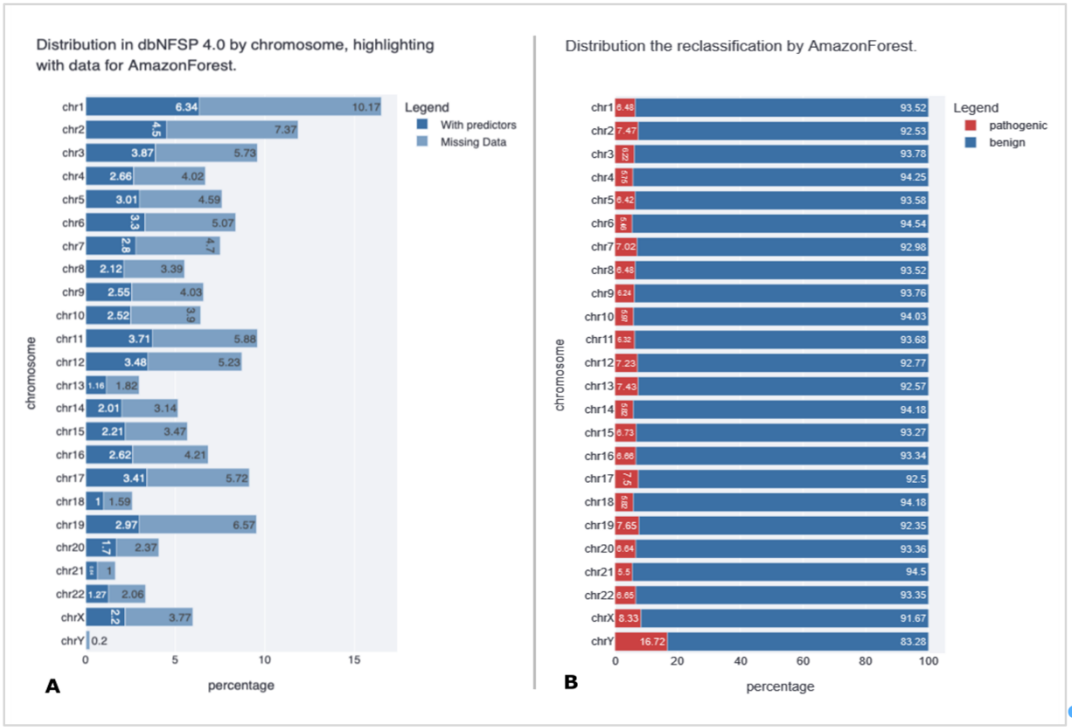



Figura 8: Proporção de variantes catalogadas no dbNFSP. Em (A), proporção de variantes anotadas funcionalmente com oito preditores de impacto. Em (B), proporção de variantes classificadas utilizando o AmazonForest (RFprob >= 0.95). Fonte: Autor.

### 4.3.1 IMPACTO FUNCIONAL EM VIAS BIOLÓGICAS

Partido do trabalho de Sanchez-Vega *et al.* (2018), analisamos a distribuição de variantes patogênicas em dez vias: ciclo celular, Hippo, Myc, Notch, Nrf2, PI-3-quinase/Akt, RTK-RAS, sinalização de TGF$\beta$, p53 e $\beta$-catenina/Wnt. Por contagem, a via que abriga mais variantes patogênicas é a NOTCH com aproximadamente 26k (32%) de variantes, seguido da via TGF com aproximadamente 6,7K (20%) de variantes genéticas catalogadas.

Observando com mais detalhe os dados na tabela 2, a quantificação da via NOCTH chega aproximadamente a 26,35%. Quando observado o percentual para os que possui frequências anotados do *EXAC*, esse índice está em 0.59% variantes, e fazendo uma conexão deste mesmo conjunto com os encontrados no 1000 Genomes é apresentado neste caso 0.06% registrados das variantes.

Tabela 2: Distribuição de variantes benignas e patogênicas em vias relacionadas a câncer e distribuição em base de dados de genômica populacional.

| Pathway | Number of variants | benign variants (%) | high probability pathogenic variants | On EXAC (%) | On 1000 Genomes Project (%) |
|---|---|---|---|---|---|
| NOTCH | 99356 | 73.65 | 26.35 | 0.59 | 0.06 |
| TGF | 32000 | 79.15 | 20.85 | 0.39 | 0.03 |
| PI3K | 36195 | 92.85 | 7.15 | 0.06 | 0 |
| WNT | 32730 | 93.22 | 6.78 | 0.15 | 0.02 |
| TP53 | 29241 | 92.65 | 7.35 | 0.32 | 0.01 |
| RTK RAS | 31674 | 93.78 | 6.22 | 0.06 | 0.01 |
| NRF2 | 13275 | 93.49 | 6.51 | 0.09 | 0 |
| HIPPO | 19812 | 96.05 | 3.95 | 0.06 | 0 |
| Cell Cycle | 12178 | 96.35 | 3.65 | 0.1 | 0 |
| MYC | 1066 | 80.21 | 19.79 | 0.66 | 0 |

### 4.3.2 MAPEAMENTO POPULACIONAL DE MUTAÇÕES SOMÁTICAS

A distribuição de variantes patogênicas em populações continentais é em sua maior parte raras, entretanto, nas nossas análises identificamos 936 variantes de acordo com as frequências calculadas a partir do *EXAC* devidamente registrados no dbNSFP para as populações. Na Figura 9 (A) podem ser observadas a distribuição de frequências nas populações Africanas, Europeias e Latino-Americanas, com um total de 815 variantes para

este grupo. Identificamos diversidade genética entre essas populações referentes a distribuição de frequências (ver Figura 9 B). Com isso, 128 variantes foram encontradas em Africanos, 544 variantes em Europeus, em Latino-Americanos 143. Apenas 25 variantes foram encontradas em todas as populações, os genes que abrigam essas variantes são: *ACVR2A*, *ATM*, *CDKN2 B*, *CREBBP*, *EP300*, *JAG1*, *JAG2*, *NOTCH1*, *NOTCH2*, *NOTCH3*, *NOTCH4*, *SMAD*/*SMAD3*/*SMAD4*. As variantes rs371059184-ACVR2A, rs79176844-JAG1 e rs200520088 *NOTCH1*, apresentam-se comuns com frequências no range de 0.1 a 0.46 entre as populações. As demais variantes são raras como mostra a Figura 9 (A). A correlação entre as frequências representadas na Figura 9 é de 0.98, assim as frequências destas variantes são quase dependentes.

Um dos resultados do estudo *in silico é apresentado* na Figura 10 (A), na qual mostra a distribuição das 936 variantes encontradas para os genes das vias estudadas e com frequências alélicas presentes *EXAC-dbNFSP*. Já na Figura 10 (B), apresentamos os resultados das contagens obtidas pelo processo de enriquecimento dos genes destas vias junto a base de dados do COSMIC, que mostram a contagem distinta por rs únicos de 84 registros nas via TP53 temos 29 rs, seguido da via PI3K com 22 rs, via NOTHC com 13 rs, RTK-RAS com 8 rs, via TGF com 6 rs e WNT com 6 rs encontrados. Por outro lado, com relação ao tipo de câncer encontrado, demos destaques para os seguintes: i) carcinoma (49) em todas as seis vias relatadas, com destaque para a via TP53 (14) e PI3K (13); ii) câncer do sistema linfático (14) nas vias TP53, NOTCH, RTK-RAS e PI3K; iii) glioma (7) nas vias PI3K e RTK-RAS; iv) neoplasias hematológicas (5) nas vias NOTCH, TP53 e RTK-RAS.

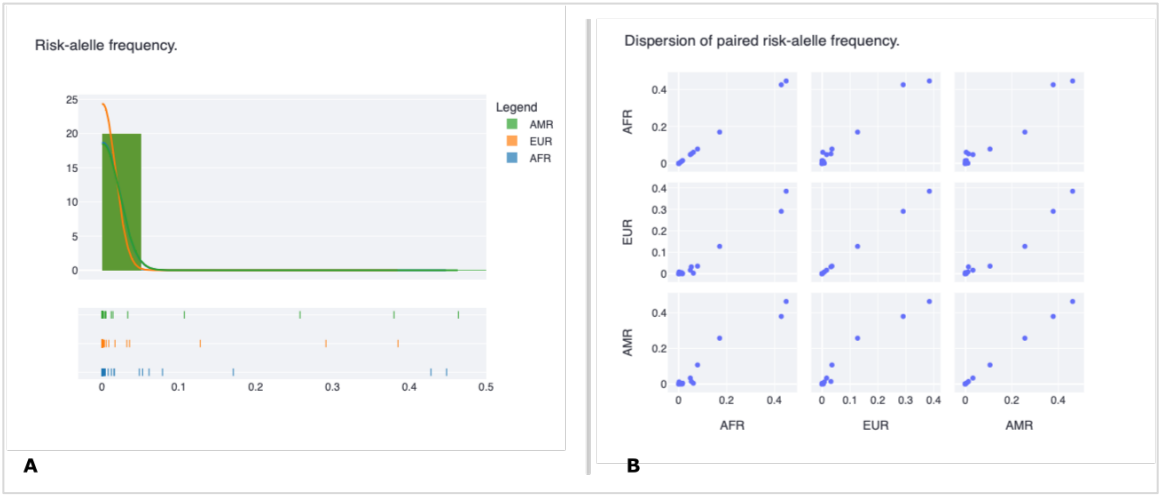



Figura 9: Informações dPOP. Em (A), frequências alélicas das variantes patogênicas com informações de frequências alélicas presentes no *dbNSFP* pelas dez vias. Em (B), dispersão de variantes patogênicas com informações de frequências alélicas presentes no *dbNSFP* pelas dez vias. Fonte: Autor.

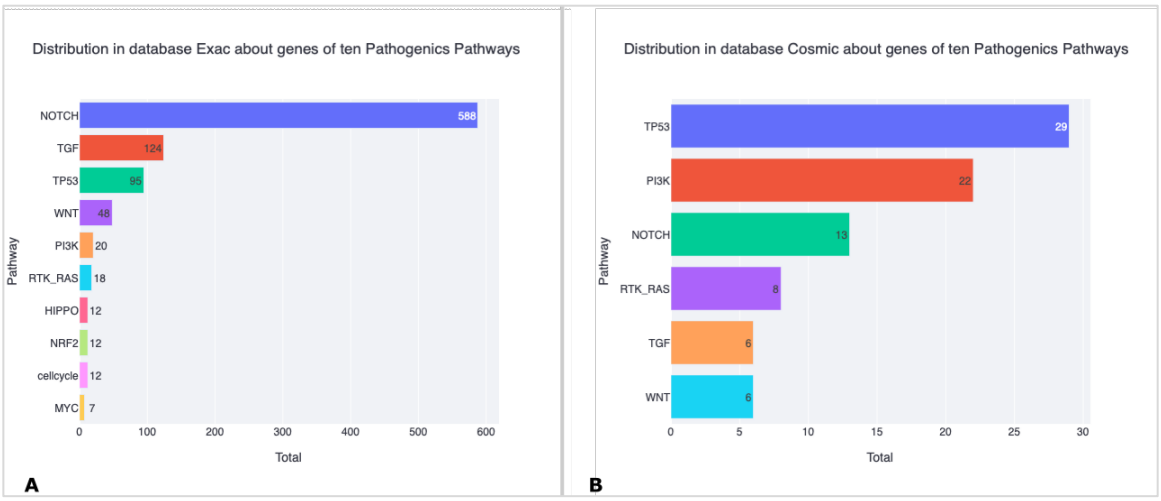



Figura 10: Informações de variantes patogênicas enriquecidas com *EXAC* e *COSMIC*. Em (A), total de variantes por vias com informações de frequências alélicas presentes no *EXAC* do *dbNSFP*. Em (B), total de variantes por vias com informações do tipo câncer presentes no *COSMIC*. Fonte: Autor.

## 4.4 DISCUSSÕES

Em estudos de genética de populações uma prática comum é analisar a patogenicidade de variantes genéticas a partir de preditores funcionais, entretanto, essa tarefa tem sido realizada de forma subjetiva, sendo que cada grupo de pesquisa estabelece regras de forma arbitrária para classificar patogenicidade com um ou mais preditores. Por exemplo, Junior, Sousa e Guerreiro (2022) define suas regras da seguinte forma: benigno, quando três ou mais preditores classificaram a variante como benigna; patogênica, quando três ou mais preditores classificaram a mutação como patogênica; inconclusivo, quando pelo menos um preditor não prediz o impacto, ou dois a classificaram como patogênica e outros dois a classificaram como benigna ou quando nenhuma predição foi feita por múltiplos preditores. A mesma tarefa foi empregada de forma similar em Reis *et al.* (2017) que aplicou preditores *in silico* sob variantes do gene HBB.

Iniciativas como o *ClinVar* e o *dbNSFP* como ferramentas bioinformáticas que agregam dados para análises funcionais não-sinônimas são relevantes para investigação clínica, além de permitir análises de genômica populacional. Entretanto, há ainda "*gaps*" de interpretação funcional de um conjunto considerável de variantes genéticas. Nesse contexto, o trabalho desenvolvido apresenta uma análise de predição global de variantes em escala genômica para consolidar as predições funcionais, que posteriormente será integrada a base de dados do AmazonForest para consulta online ou via programação em linguagem R.

Da aplicação global, investigamos variantes em vias biológicas relacionadas ao câncer. Notamos que o estudo *in silico* aplicado na investigação das vias conseguiu mostrar informações adicionais sobre as mesmas, realizando um enriquecimento sobre os tipos de câncer pelos alvos de procura do estudo.

Em estudos de genômica populacional as ferramentas bioinformáticas que auxiliam na predição e anotação de variantes genéticas são de extrema importância. O presente estudo

apresenta o desenvolvimento e expansão da base de dados do *AmazonForest* para nível genômico e diferenças de frequências alélicas em variantes patogênicas considerando as populações Europeias, Africanas e Latino-Americanas. O estudo aprofundado dessas variantes considerando a África e a Europa pode subsidiar estudos de câncer nas populações miscigenadas como a população brasileira, que tem sua origem parental tri-hibrida (África, Europa e Nativa Americana). A genética populacional e as abordagens bioinformáticas integrativas são elementos essenciais para construir o conhecimento a partir de grandes volumes de dados biológicos heterogêneos, uma vez que essas abordagens têm permitido identificar novos marcadores relacionados aos processos cancerígenos.

A análise sistemática de dados biológicos é fundamental para a prática de investigação, diagnóstico e planejamento de terapias para diversos tipos de doenças complexas. No entanto, a complexidade, alta dimensionalidade e heterogeneidade dos dados permanecem um obstáculo significante para esta análise. Isso requer a construção de esquemas de dados integrados, agregando e visualizando dados biológicos e aplicando métodos computacionais e estatísticos mais robustos para elucidar a base molecular dos fenótipos.

## 4.5 CONCLUSÕES E TRABALHOS FUTUROS

No decorrer do trabalho foi demonstrado, por meio de estudos de referência, a importância da realização de estudos *in silico* na pesquisa de doenças. O que neste estudo foi realizado por meio da aplicação de um modelo de meta-predição, e na demonstração de sua utilidade para o desenvolvimento e descoberta em pesquisas de doenças complexas como o câncer.

Percorremos um caminho que foi desde a mineração e utilização de bases de dados públicas, como a aplicação do processo de transformações e reclassificações com algoritmos

meta-preditivos. Além de buscar enriquecimentos dos achados com base de dados de informações sobre a doença.

A nova geração de dados com aproximadamente 52 milhões de registros é mais um outro importante artefato gerado por esta etapa do estudo. Após todo processo de pré-tratamento e reclassificação com AmazonForest, conseguimos de forma surpreendente um dataset tão expressivo e rico em informações, que pelos módulos iniciais desenvolvidos, já conseguiu mostrar sua aplicabilidade, como no estudo inicial das dez vias importantes do câncer.

Percebemos que ainda é possível desbravar mais conhecimento com este novo dataset, pois o mesmo ainda possui outras informações catalogadas que podem ser utilizadas, transformadas e completadas com novas perspectivas e interpretações possíveis de gerar novas funcionalidades.

Todo processo para estabelecer a classificação de patogenicidade de uma variante é complexo. Quando se trata de estudos em etapas humanas, as interpretações de um especialista são fundamentais para uma conclusão mais acertada. A utilização de softwares *in silico* entra nesse auxílio para fortalecer a qualidade da interpretação. Além de utilizar outros recursos de associação com diversos bancos de dados, frequência alélica na população e histórico clínico familiar.

A devida integração com o AmazonForest e a devida adequação de sua interface para consultas a este novo dataset, devem ser implementadas e, futuramente, estará disponível para comunidade científica.

# 5 CONSIDERAÇÕES FINAIS

Acredita-se que o estudo do perfil genômico pode se tornar um padrão para o atendimento em oncologia clínica. Mas para que isso ocorra, compreendemos ser necessário que haja grande compartilhamento de dados para o enriquecimento dos conhecimentos e levem ao progresso da medicina de precisão.

Exemplo de estudos em bases públicas, como as que utilizam o *TGCA* (*The Cancer Genome Atlas Program*), conseguem fazer a análise de alterações estabelecidas clinicamente relevantes, destacando as alterações moleculares para as quais os testes são atualmente recomendados e ajudam a oportunizar a expansão das indicações para o uso de terapias direcionadas. De uma outra forma, tentam apresentar novos recurso para investigação, validação da relevância clínica e identificação para novos alvos terapêuticos.

Dentro deste contexto, o presente trabalho buscou em um primeiro momento melhorar as informações contidas no banco de dados *ClinVar*, principalmente para tentar obter uma visão mais assertiva sobre o conjunto de variantes que se encontram como VUS e com conflito de interpretação. Para conseguir esse objetivo utilizou-se de técnicas de *machine learn*, para se chegar a esses resultados iniciais.

Com o conhecimento obtido dentre aplicações de vários modelos e pela produção de um framework para aplicações dos modelos, evoluímos para um produto final que compilou nossas expectativas - o *AmazonForest* - que permite realizar a investigação de variantes gênicas, assim como possibilita a caracterização para aquelas com alta tendência para patogenicidade, através da aplicação de um modelo treinado por *random forest* de forma simplificada e eficiente.

O novo conjunto de dados obtido pelo estudo torna-se importante componente a ser utilizado para novas pesquisas. Sabemos que essa jornada apenas iniciou, pois as possibilidades de

integração, enriquecimentos e desafios ainda são diversos e ilimitados. Esse estudo mostrou apenas um prisma do quanto pode ser alcançado nessa busca por novos conhecimentos.

# 6 REFERÊNCIAS


ALMARRI, M. A. *et al.* Population structure, stratification, and introgression of human structural variation. **Cell**, Elsevier, 2020.

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2020.

BAHCECI, I. *et al.* Pathwaymapper: a collaborative visual web editor for cancer pathways and genomic data. **Bioinformatics**, Oxford University Press, v. 33, n. 14, p. 2238–2240, 2017.

BATMANIAN, L.; RIDGE, J.; WORRALL, S. **Biochemistry for health professionals**. [S.l.]: Elsevier Australia, 2011.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1798–1828, 2013.

BENSON, D. A. *et al.* Genbank. **Nucleic acids research**, Oxford University Press, v. 41, n. D1, p. D36–D42, 2012.

BESSIERE, P. *et al.* **Bayesian programming**. [S.l.]: CRC press, 2013.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the fifth annual workshop on Computational learning theory**. [S.l.: s.n.], 1992. p. 144–152.

BOSIO, M. *et al.* ediva—classification and prioritization of pathogenic variants for clinical diagnostics. **Human mutation**, Wiley Online Library, v. 40, n. 7, p. 865–878, 2019.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

CAMPBELL, J. D. *et al.* Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. **Cell reports,** Elsevier, v. 23, n. 1, p. 194–212, 2018.

CERAMI, E. *et al.* The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. **Cancer discovery**, AACR, v. 2, n. 5, p. 401–404, 2012.

CHAKRAVARTY, D. *et al.* Oncokb: a precision oncology knowledge base. **JCO precision oncology**, American Society of Clinical Oncology, v. 1, p. 1–16, 2017.

COATES, A.; NG, A.; LEE, H. An analysis of single-layer networks in unsupervised feature learning. In: JMLR WORKSHOP AND CONFERENCE PROCEEDINGS. **Proceedings of the fourteenth international conference on artificial intelligence and statistics**. [S.l.], 2011. p. 215–223.

CONSORTIUM, . G. P. *et al.* A global reference for human genetic variation. **Nature**, Nature Publishing Group, v. 526, n. 7571, p. 68, 2015.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 157–175.

DICTIONARY, O. E. **Retail**. 2019.

FEUK, L.; CARSON, A. R.; SCHERER, S. W. Structural variation in the human genome. **Nature Reviews Genetics**, Nature Publishing Group, v. 7, n. 2, p. 85–97, 2006.

FREEDMAN, M. L. *et al.* Principles for the post-gwas functional characterization of cancer risk loci. **Nature genetics**, Nature Publishing Group, v. 43, n. 6, p. 513–518, 2011.

FRENCH, J.; EDWARDS, S. The role of noncoding variants in heritable disease. **Trends in Genetics**, Elsevier, 2020.

GAO, G. F. *et al.* Before and after: comparison of legacy and harmonized tcga genomic data commons' data. **Cell systems**, Elsevier, v. 9, n. 1, p. 24–34, 2019.

GOODMAN, N. Biological data becomes computer literate: new advances in bioinformatics. **Current opinion in biotechnology**, Elsevier, v. 13, n. 1, p. 68–71, 2002.

GOODWIN, W.; GOODWIN. **Forensic DNA typing protocols**. [S.l.]: Springer, 2016.

HAGEN, J. B. The origins of bioinformatics. **Nature Reviews Genetics**, Nature Publishing Group, v. 1, n. 3, p. 231–236, 2000.

HAO, J.; HO, T. K. Machine learning made easy: a review of scikit-learn package in python programming language. **Journal of Educational and Behavioral Statistics**, SAGE Publications Sage CA: Los Angeles, CA, v. 44, n. 3, p. 348–361, 2019.

HARTL, D. L.; CLARK, A. G. **Princípios de Genética de Populações-4**. [S.l.]: Artmed Editora, 2010.

HASPEL, R. L. *et al.* Teaching genomic pathology: translating team-based learning to a virtual environment using computer-based simulation. **Archives of pathology & laboratory medicine**, the College of American Pathologists, v. 143, n. 4, p. 513–517, 2019.

HAYHURST, J. *et al.* A community driven gwas summary statistics standard. **bioRxiv**, Cold Spring Harbor Laboratory, 2022.

HOGEWEG, P. The roots of bioinformatics in theoretical biology. **PLoS Comput Biol**, Public Library of Science, v. 7, n. 3, p. e1002021, 2011.

HOUDAYER, C. *et al.* Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on brca1 and brca2 variants. **Human mutation**, Wiley Online Library, v. 33, n. 8, p. 1228–1238, 2012.

HSU, C.-W. *et al.* **A practical guide to support vector classification**. [S.l.]: Taipei, 2003.

INNES, M. *et al.* On machine learning and programming languages. In: ASSOCIATION FOR COMPUTING MACHINERY (ACM). [S.l.], 2018.

JONSON, P. H.; PETERSEN, S. B. A critical view on conservative mutations. **Protein engineering**, Oxford University Press, v. 14, n. 6, p. 397–402, 2001.

JUNIOR, M. L. F. S.; SOUSA, J. V. de; GUERREIRO, J. F. Analysis of coding variants in the human fto gene from the gnomad database. **PloS one**, Public Library of Science San Francisco, CA USA, v. 17, n. 1, p. e0248610, 2022.

KULSKI, J. K. Next-generation sequencing—an overview of the history, tools, and "omic" applications. **Next Generation Sequencing–Advances, Applications and Challenges**, InTech, Rijeka, Croatia, p. 3–60, 2016.

LANDER, E. S. *et al.* Initial sequencing and analysis of the human genome. Macmillan Publishers Ltd., 2001.

LANDRUM, M. J. *et al.* Clinvar: public archive of relationships among sequence variation and human phenotype. **Nucleic acids research,** Oxford University Press, v. 42, n. D1, p. D980–D985, 2013.

LAWRENCE, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. **Nature**, Nature Publishing Group, v. 505, n. 7484, p. 495–501, 2014.

LIN, H.-T.; LIN, C.-J. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. **submitted to Neural Computation**, v. 3, n. 1-32, p. 16, 2003.

LIN, Y.; JEON, Y. Random forests and adaptive nearest neighbors. **Journal of the American Statistical Association**, Taylor & Francis, v. 101, n. 474, p. 578–590, 2006.

LIU, X. *et al.* dbnsfp v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. **Genome medicine**, BioMed Central, v. 12, n. 1, p. 1–8, 2020.

LOGUE, M. W. *et al.* A comprehensive genetic association study of alzheimer disease in african americans. **Archives of neurology**, American Medical Association, v. 68, n. 12, p. 1569–1579, 2011.

MACARTHUR, J. *et al.* The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). **Nucleic acids research**, Oxford University Press, v. 45, n. D1, p. D896–D901, 2017.

MANDOIU, I.; ZELIKOVSKY, A. **Bioinformatics algorithms: techniques and applications**. [S.l.]: John Wiley & Sons, 2008. v. 3.

MARDIS, E. R. Next-generation dna sequencing methods. **Annu. Rev. Genomics Hum**. Genet., Annual Reviews, v. 9, p. 387–402, 2008.

MARDIS, E. R. Next-generation sequencing platforms. **Annu Rev Anal Chem**, v. 6, n. 1, p. 287–303, 2013.

MERMEL, C. H. *et al.* Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. **Genome biology**, Springer, v. 12, n. 4, p. 1–14, 2011.

MINA, M. *et al.* Conditional selection of genomic alterations dictates cancer evolution and oncogenic dependencies. **Cancer cell**, Elsevier, v. 32, n. 2, p. 155–168, 2017.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

MONTINARO, F.; CAPELLI, C. A worldwide map of human structural variants. **Trends in Genetics**, Elsevier, v. 36, n. 10, p. 722–725, 2020.

MORAIS, R. M. On the suitability, requisites, and challenges of machine learning. **IEEE/OSA Journal of Optical Communications and Networking**, IEEE, v. 13, n. 1, p. A1–A12, 2020.

NAJ, A. C. *et al.* Effects of multiple genetic loci on age at onset in late-onset 74ílico74er disease: a genome-wide association study. **JAMA neurology**, American Medical Association, v. 71, n. 11, p. 1394–1404, 2014.

NURK, S. *et al.* The complete sequence of a human genome. **Science**, American Association for the Advancement of Science, v. 376, n. 6588, p. 44–53, 2022.

NWANGANGA, F.; CHAPPLE, M. **Practical machine learning in R**. [S.l.]: John Wiley & Sons, 2020.

PALHETA, H. G. A. *et al.* Amazonforest: In 74ílico metaprediction of pathogenic variants. **Biology**, MDPI, v. 11, n. 4, p. 538, 2022.

PHAN, L. Snps classification and terminology: dbsnp reference snp (rs) gene and consequence annotation. In: **Single Nucleotide Polymorphisms**. [S.l.]: Springer, 2022. P. 3–12.

REIS, T. Carlice-dos *et al.* Investigation of mutations in the hbb gene using the 1,000 genomes database. **PloS One**, Public Library of Science San Francisco, CA USA, v. 12, n. 4, p. e0174637, 2017.

RISH, I. *et al.* Na empirical study of the naive bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. [S.l.: s.n.], 2001. V. 3, n. 22, p. 41–46.

RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 2013.

RUWALD, M. H. *et al.* Stop-codon and c-terminal nonsense mutations are associated with a lower risk of cardiac events in patients with long qt syndrome type 1. **Heart Rhythm**, Elsevier, v. 13, n. 1, p. 122–131, 2016.

SANCHEZ-VEGA, F. *et al.* Oncogenic signaling pathways in the cancer genome atlas. **Cell**, Elsevier, v. 173, n. 2, p. 321–337, 2018.

SANTOS, A. M. Ribeiro-dos. **DIVERSIDADE DE INDELS EM SÍTIOS DE LIGAÇÃO DE FATORES DE TRANSCRIÇÃO HUMANOS**. Dissertação (Mestrado) — Universidade Federal do Pará, 9 2014.

SANTOS, A. M. Ribeiro-dos *et al.* High-throughput sequencing of a south american amerindian. **PLoS One**, Public Library of Science San Francisco, USA, v. 8, n. 12, p. e83340, 2013.

SAYERS, E. W. *et al.* Database resources of the national center for biotechnology information. **Nucleic acids research**, Oxford University Press, v. 47, n. Database issue, p. D23, 2019.

SAYERS, E. W. *et al.* Database resources of the national center for biotechnology information. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D13–D25, 2012.

SHERMAN, R. M.; SALZBERG, S. L. Pan-genomics in the human genome era. Nature **Reviews Genetics**, Nature Publishing Group, p. 1–12, 2020.

SHYR, C.; KUSHNIRUK, A.; WASSERMAN, W. W. Usability study of clinical exome analysis software: top lessons learned and recommendations. **Journal of biomedical informatics**, Elsevier, v. 51, p. 129–136, 2014.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Database system concepts**. [S.l.]: McGraw-Hill, 2011.

SINGH, P. K.; MISTRY, K. N. A computational approach to determine susceptibility to cancer by evaluating the deleterious effect of nssnp in xrcc1 gene on binding interaction of xrcc1 protein with ligase iii. **Gene**, Elsevier, v. 576, n. 1, p. 141–149, 2016.

STROBL, C. *et al.* Bias in random forest variable importance measures: Illustrations, sources and a solution. **BMC bioinformatics**, Springer, v. 8, n. 1, p. 25, 2007.

SUDMANT, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. **Nature**, Nature Publishing Group, v. 526, n. 7571, p. 75–81, 2015.

TIMPSON, N. J. *et al.* Genetic architecture: the shape of the genetic contribution to human traits and disease. **Nature Reviews Genetics**, Nature Publishing Group, v. 19, n. 2, p. 110, 2018.

VENTER, J. C. *et al.* The sequence of the human genome. **science**, American Association for the Advancement of Science, v. 291, n. 5507, p. 1304–1351, 2001.

VISSCHER, P. M. *et al.* 10 years of gwas discovery: biology, function, and translation. **The American Journal of Human Genetics**, Elsevier, v. 101, n. 1, p. 5–22, 2017.

XUE, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. **Nature communications**, Nature Publishing Group, v. 9, n. 1, p. 1–14, 2018.

# APÊNDICE A – ATIVIDADES

## A.1 Publicações Científicas

### A.1.1 Publicação de artigos

1. PALHETA, H.G.A., H., Gonçalves, W. G., Brito, L. M., Santos, A. R., Matsumoto, M., Ribeiro-dos-Santos, Â., & Araújo, G. S. *AmazonForest: In-silico meta-prediction of pathogenic variants.* Biology, MDPI, v. 11,n.4, p.538, 2022.

2. GONÇALVES, Wanderson Gonçalves , Santos, M. H. P. D., Brito, L. M., PALHETA, H.G.A., Lobato, F. M. F., Demachki, S., ... & Araújo, G. S. D. *DeepHP: A New Gastric Mucosa Histopathology Dataset for Helicobacter pylori Infection Diagnosis*. International Journal of Molecular Sciences, v. 23, n. 23, p. 14581, 2022.

3. FAZZI-GOMES, Paola, P., Aguiar, J., Cabral, G. F., Marques, D., PALHETA, H.G.A., Moreira, F., ... & Santos, S.. *Genomic approach for conservation and the sustainable management of endangered species of the Amazon.* Plos one, v. 16, n. 2, p. e0240002, 2021.

biology

MDPI

*Article*

# AmazonForest: In Silico Metaprediction of Pathogenic Variants

**Helber Gonzales Almeida Palheta [1], Wanderson Gonçalves Gonçalves [1,2], Leonardo Miranda Brito [1], Arthur Ribeiro dos Santos [1], Marlon dos Reis Matsumoto [1], Ândrea Ribeiro-dos-Santos [1,2,†] and Gilderlanio Santana de Araújo [1,*,†]**

[1] Laboratory of Human and Medical Genetics, Graduate Program of Genetics and Molecular Biology, Institute of Biological Sciences, Federal University of Pará, Belém 66075-110, Brazil; hpalheta@gmail.com (H.G.A.P.); wandersongegoncalves@gmail.com (W.G.G.); lb9458@gmail.com (L.M.B.); arthurrdsantos@outlook.com (A.R.d.S.); marlonmatsumotosdb@gmail.com (M.d.R.M.); akelyufpa@gmail.com (Â.R.-d.-S.)
[2] Research Center on Oncology, Graduate Program of Oncology and Medical Science, Federal University of Pará, Belém 66073-000, Brazil
* Correspondence: gilderlanio@gmail.com
† These authors contributed equally to this work.

**Simple Summary:** ClinVar is a valuable platform that stores a large set of relevant genetic associations with complex phenotypes. However, the functional impact of a partial set of such associations remains misinterpreted, due to the presence of variants with uncertain significance or with conflicting pathogenicity interpretations. To fill this gap, we present AmazonForest: a metaprediction model based on Random Forest for pathogenicity prediction. AmazonForest was used to reclassify a set of ∼101,000 variants that were predicted as having high pathogenic probability. AmazonForest is available as a web tool with a simple web interface, and also as an R object for pathogenicity predictions.

**Abstract:** ClinVar is a web platform that stores ∼789,000 genetic associations with complex diseases. A partial set of these cataloged genetic associations has challenged clinicians and geneticists, often leading to conflicting interpretations or uncertain clinical impact significance. In this study, we addressed the (re)classification of genetic variants by AmazonForest, which is a random-forest-based pathogenicity metaprediction model that works by combining functional impact data from eight prediction tools. We evaluated the performance of representation learning algorithms such as autoencoders to propose a better strategy. All metaprediction models were trained with ClinVar data, and genetic variants were annotated with eight functional impact predictors cataloged with SnpEff/SnpSift. AmazonForest implements the best random forest model with a one hot data-encoding strategy, which shows an Area Under ROC Curve of ≥0.93. AmazonForest was employed for pathogenicity prediction of a set of ∼101,000 genetic variants of uncertain significance or conflict of interpretation. Our findings revealed ∼24,000 variants with high pathogenic probability ($RF_{prob} \geq 0.9$). In addition, we show results for Alzheimer's Disease as a demonstration of its application in clinical interpretation of genetic variants in complex diseases. Lastly, AmazonForest is available as a web tool and R object that can be loaded to perform pathogenicity predictions.

**Keywords:** metaprediction; encoding data; random forest; representation learning; genetic variants; clinical impact; functional impact



**Citation:** Palheta, H.G.A.; Gonçalves, W.G.; Brito, L.M.; dos Santos, A.R.; dos Reis Matsumoto, M.; Ribeiro-dos-Santos, Â.; de Araújo, G.S. AmazonForest: In Silico Metaprediction of Pathogenic Variants. *Biology* **2022**, *11*, 538. https://doi.org/10.3390/biology11040538

Academic Editor: Wojciech Makalowski

Received: 26 January 2022
Accepted: 2 March 2022
Published: 31 March 2022



## 1. Introduction

Next-generation sequencing (NGS) methods have allowed whole-genome analyses for humans and other species. Genome-wide association studies (GWAS) and candidate gene studies have produced a large volume of genetic associations between single-nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) with complex diseases. Most of these associations show variable effects and genetic diversity among populations [1,2].

International Journal of *Molecular Sciences*

MDPI

*Article*

# DeepHP: A New Gastric Mucosa Histopathology Dataset for *Helicobacter pylori* Infection Diagnosis

**Wanderson Gonçalves e Gonçalves [1,2], Marcelo Henrique Paula dos Santos [1], Leonardo Miranda Brito [1], Helber Gonzales Almeida Palheta [1], Fábio Manoel França Lobato [3], Samia Demachki [1], Ândrea Ribeiro-dos-Santos [1,2,*] and Gilderlanio Santana de Araújo [1,*]**

[1] Laboratory of Human and Medical Genetics, Institute of Biological Sciences, Graduate Program of Genetics and Molecular Biology, Federal University of Pará, Belém 66075-110, Brazil
[2] Research Center on Oncology, Graduate Program of Oncology and Medical Science, Federal University of Pará, Belém 66073-000, Brazil
[3] Laboratory of Applied Computing, Engineering and Geoscience Institute, Federal University of Western Pará, Santarém 68040-255, Brazil
* Correspondence: akelyufpa@gmail.com (Â.R.-d.-S.); gilderlanio@gmail.com (G.S.d.A.)

**Abstract:** Emerging deep learning-based applications in precision medicine include computational histopathological analysis. However, there is a lack of the required training image datasets to generate classification and detection models. This phenomenon occurs mainly due to human factors that make it difficult to obtain well-annotated data. The present study provides a curated public collection of histopathological images (DeepHP) and a convolutional neural network model for diagnosing gastritis. Images from gastric biopsy histopathological exams were used to investigate the performance of the proposed model in detecting gastric mucosa with *Helicobacter pylori* infection. The DeepHP database comprises 394,926 histopathological images, of which 111 K were labeled as *Helicobacter pylori* positive and 283 K were *Helicobacter pylori* negative. We investigated the classification performance of three Convolutional Neural Network architectures. The models were tested and validated with two distinct image sets of 15% (59K patches) chosen randomly. The VGG16 architecture showed the best results with an Area Under the Curve of 0.998%. The results showed that CNN could be used to classify histopathological images from gastric mucosa with marked precision. Our model evidenced high potential and application in the computational pathology field.

**Keywords:** deep learning; *Helicobacter pylori*; computational pathology



**Citation:** Gonçalves, W.G.e.; Santos, M.H.P.d.; Brito, L.M.; Palheta, H.G.A.; Lobato, F.M.F.; Demachki, S.; Ribeiro-dos-Santos, Â.; Araújo, G.S.d. DeepHP: A New Gastric Mucosa Histopathology Dataset for *Helicobacter pylori* Infection Diagnosis. *Int. J. Mol. Sci.* **2022**, *23*, 14581. https://doi.org/10.3390/ijms232314581

Academic Editor: Alexandre G. de Brevern

Received: 24 September 2022
Accepted: 16 November 2022
Published: 23 November 2022



## 1. Introduction

Gastric cancer is the fifth most common cancer worldwide and the fourth in deaths caused by cancer in 2020 [1]. *Helicobacter pylori* (HP) infection is the main risk factor accounting for around 89% of the distal gastric cancer cases around the world [2]. The HP is extremely adapted to the human gastric mucosa. However, disordered HP proliferation induces inflammation in the gastric mucosa, which can sequentially result in gastric cancer development [3]. It can bind to epithelial cells and prevent the immune response to cancer. HP has a high prevalence worldwide, of almost 50% [4], and 1–3% of the cases of *H. pylori* infection progress to gastric cancer [5].

Endoscopy is the main procedure for assessing HP infection and gastric cancer, followed by an histopathological biopsy analysis [6]. The histopathological analysis allows for identifying HP ubiquity and morphological alterations in the gastric mucosa. A gastric biopsy analysis is a highly time-consuming task. It can be affected by biotechnological factors such as staining techniques, errors in gathering biopsy sites, and also the pathologists' subjectivity/experience [7,8].

Intelligent computational models are promising in the medical domain in terms of assisting clinical decisions [9]. Few studies have been proposed for gastric-related dis-

PLOS ONE

RESEARCH ARTICLE

# Genomic approach for conservation and the sustainable management of endangered species of the Amazon

**Paola Fazzi-Gomes[1], Jonas Aguiar[2], Gleyce Fonseca Cabral[1], Diego Marques[1], Helber Palheta[1], Fabiano Moreira[3], Marilia Rodrigues[4], Renata Cavalcante[5], Jorge Souza[5], Caio Silva[1], Igor Hamoy[4], Sidney Santos[1,3,6]***

**1** Human and Medical Genetics Lab, Biological Sciences Institute, Federal University of Pará, Belem, PA, Brazil, **2** Federal University of Pará –Campus Bragança, Alameda Leandro Ribeiro s/n, Bragança, PA, Brazil, **3** Graduate Program in Oncology and Medical Sciences, Center of Oncology Research, Federal University of Pará, Belém, Pará, Brazil, **4** Rural Federal University of the Amazon, Applied Genetics Lab, Socioenvironmental and Water Resources Institute, Belem, PA, Brazil, **5** Bioinformatics Graduate Program, Metropole Digital Institute, Federal University of Rio Grande do Norte, Natal, RN, Brazil, **6** Graduate Program in Genetics and Molecular Biology, Laboratory of Human and Medical Genetics, Federal University of Pará, Belém, Pará, Brazil

* sidneysantos@ufpa.br



OPEN ACCESS

**Citation:** Fazzi-Gomes P, Aguiar J, Cabral GF, Marques D, Palheta H, Moreira F, et al. (2021) Genomic approach for conservation and the sustainable management of endangered species of the Amazon. PLoS ONE 16(2): e0240002. https://doi.org/10.1371/journal.pone.0240002

**Editor:** Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

**Received:** September 15, 2020

**Accepted:** December 10, 2020

**Published:** February 24, 2021



**Data Availability Statement:** The development of this study was based on the complete genome of *Arapaima gigas*, published by Vialle et al., 2018, available from the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/assembly), ID: 12404. Raw sequencing data are available at Figshare Data Repository: https://doi.org/10.6084/m9.figshare.8088533, https://doi.org/10.6084/m9.figshare.8088629.

## Abstract

A broad panel of potentially amplifiable microsatellite loci and a multiplex system were developed for the Amazonian symbol fish species Arapaima gigas, which is currently in high danger of extinction due to the disorderly fishing exploitation. Several factors have contributed to the increase of this threat, among which we highlight the lack of genetic information about the structure and taxonomic status of the species, as well as the lack of accurate tools for evaluation of the effectivity of current management programs. Based on Arapaima gigas' whole genome, available at the NCBI database (ID: 12404), a total of 95,098 unique perfect microsatellites were identified, including their proposed primers. From this panel, a multiplex system containing 12 tetranucleotide microsatellite markers was validated. These tools are valuable for research in as many areas as bioinformatics, ecology, genetics, evolution and comparative studies, since they are able to provide more accurate information for fishing management, conservation of wild populations and genetic management of aquaculture.

## 1. Introduction

The species *Arapaima gigas* (Schinz, 1822) belongs to the Arapaimidae family–order of the Osteoglossiformes [1], which composes one of the oldest groups of teleost fishes. It is the world's largest scale fish, and specimens may reach up to 200 kg of body mass and 3 m of length [2]. The species can be found in basins of South American countries, such as Brazil, Peru, Colombia, Ecuador, Bolivia and Guyana [3, 4]. Since years ago, *A. gigas* has been relevant in aquaculture due to its fast growth, high fillet yield, mild-flavored white meat, and great market acceptance, both domestically and abroad [5, 6].

A.1.2 Publicação em Anais de Evento

1. PALHETA, H.G.A.; RIBEIRO-DOS-SANTOS, A.; ARAÚJO, G.; Kurtz, G., S.; *BRAVA: a tool for exploring clinical impact variants in Native Americans.*. In: 2nd International Meeting on Oncology Research, Belém, 2019.

2. SILVA, M. N. S.; RIBEIRO-DOS-SANTOS, A; SCHAAN, A. P.; PALHETA, H.G.A.;SARQUIS, D.; CARLICE-DOS-REIS, T.; *Investigation of variants in genes involved with hereditary cancer syndromes at the 1000 Genomes database.*. In: IV Simpósio NorteNordeste de Bioinformática, Bragança, 2019.

3. GOMES, P. F. F.; AGUIAR, J. P.; PALHETA, H.G.A.; MOREIRA, F., C.; RODRIGUES, M., D.; CAVALCANTE, R.; SOUZA, J., E.; HAMOY, I., G.; RIBEIRO-DOS-SANTOS, S.; *Identification of microsatellite markers in Arapaima gigas' genome (Schinz, 1822).*. In: IV Simpósio Norte-Nordeste de Bioinformática, Bragança, 2019.

4. PALHETA, H.G.A.; GONÇALVES, W.; RIBEIRO-DOS-SANTOS, A; ARAÚJO, G; AmazonForest: *A Meta-Prediction Model for Reclassification of Genetic Variants*. In: SIG 2021: Machine Learning in Bioinformatics. RSG- Brazil, 2021.

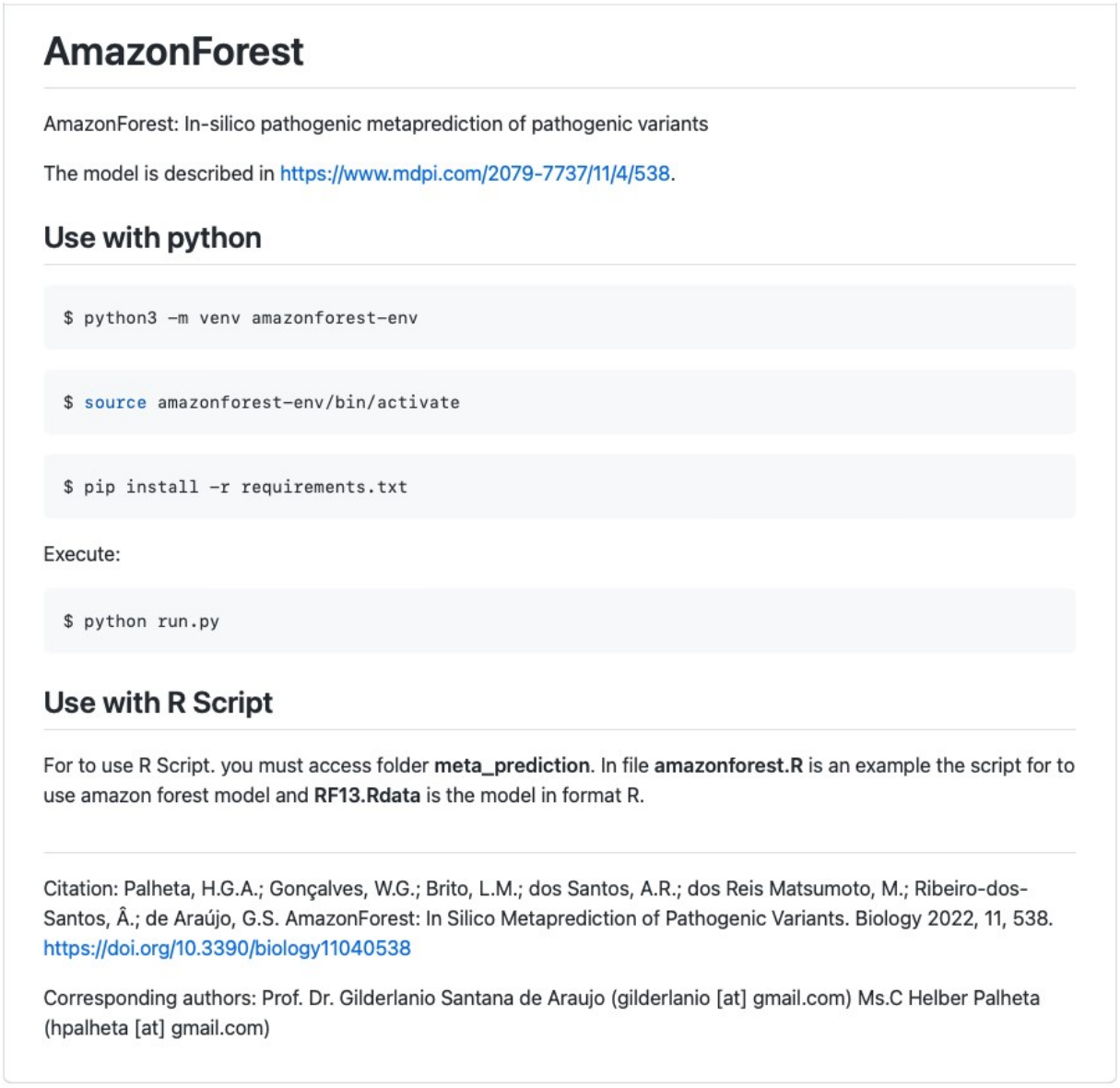
5. GONÇALVES, WANDERSON GONÇALVES E; BRITO, L. M., PALHETA, H.G.A.; SANTOS, M.; LOBATO, F.; RIBEIRO-DOS-SANTOS, A.; ARAUJO, G. *Automated Classification of Inflammatory Profiles by Convolutional Neural Network*. In: SIG 2021: Machine Learning in Bioinformatics. RSG- Brazil, 2021.

6. PALHETA, H.G.A.; RIBEIRO-DOS-SANTOS, A; ARAÚJO, G.; *Mining genome-wide data to consolidate pathogenic predictions: cancer applications*. In: 3nd International Meeting on Oncology Research, Belém, 2022.

# APÊNDICE B – COMPLEMENTARES AO AMAZONFOREST

## B.1- AmazonForest no GitHub

Disponibilizamos uma versão para acompanhamento do AmazonForest no GitHub.



# AmazonForest

AmazonForest: In-silico pathogenic metaprediction of pathogenic variants

The model is described in https://www.mdpi.com/2079-7737/11/4/538.

## Use with python

```
$ python3 -m venv amazonforest-env
```

```
$ source amazonforest-env/bin/activate
```

```
$ pip install -r requirements.txt
```

Execute:

```
$ python run.py
```

## Use with R Script

For to use R Script. you must access folder **meta_prediction**. In file **amazonforest.R** is an example the script for to use amazon forest model and **RF13.Rdata** is the model in format R.

---

Citation: Palheta, H.G.A.; Gonçalves, W.G.; Brito, L.M.; dos Santos, A.R.; dos Reis Matsumoto, M.; Ribeiro-dos-Santos, Â.; de Araújo, G.S. AmazonForest: In Silico Metaprediction of Pathogenic Variants. Biology 2022, 11, 538. https://doi.org/10.3390/biology11040538

Corresponding authors: Prof. Dr. Gilderlanio Santana de Araujo (gilderlanio [at] gmail.com) Ms.C Helber Palheta (hpalheta [at] gmail.com)

Figura 11: Home page do Amazonforest no github.

B.2 Script em linguagem R par utilização do AmazonFoest

Para utilizar o script em linguagem R é necessário fazer o download dos arquivos RData em no GitHub do projeto:
https://github.com/hpalheta/amazonforest/tree/master/meta_prediction

```
#!/usr/bin/env Rscript
#'
#' This is the R Script to perform AmazonForest for functional impact of variant genetics.
#'
#' The AmazonForest model is a random forest based model trained and tested with eight single
#' functional predictors
#'
#' The model is described in https://www.mdpi.com/2079-7737/11/4/538.
#'
#' Citation: Palheta, H.G.A.; Gonalves, W.G.; Brito, L.M.; dos Santos, A.R.; dos Reis Matsumoto, M.;
#'          Ribeiro-dos-Santos,    .; de Araujo, G.S. AmazonForest: In Silico Metaprediction of Pathogenic
#'          Variants. Biology 2022, 11, 538. https://doi.org/10.3390/biology11040538
#'
#' Corresponding authors: Prof. Dr. Gilderlanio Santana de Araujo (gilderlanio [at] gmail.com)
#'                        Ms.C Helber Palheta (hpalheta [at] gmail.com)



#' 0) Required libraries.

suppressPackageStartupMessages(library("ROCR"))
suppressPackageStartupMessages(library("randomForest"))
suppressPackageStartupMessages(library("caret"))


#' 1) The input comprise data for eight functional predictors. We extracted data from SNPSift data for example.
#' Below, a simple way of input to perform a new prediction.

FATHMM <- 'D'
LRT_pred <- 'N'
MutaAss <- 'L'
MutaTaster <- 'N'
PROVEAN <- 'D'
Pph2_HDIV <- 'D'
Pph2_HVAR <- 'P'
SIFT <- 'D'

predict_values <- c(FATHMM, LRT_pred, MutaAss, MutaTaster, PROVEAN, Pph2_HDIV, Pph2_HVAR, SIFT)
to_predict <- data.frame(FATHMM, LRT_pred, MutaAss, MutaTaster, PROVEAN, Pph2_HDIV, Pph2_HVAR, SIFT)


#'2) Load categorial data for CinVar stored variants. All variantes were annotated for 8 functional predictors.

dataset_original <- read.csv(pathdata)

```

```r
dataset_original <- transform(
  dataset_original,
  CLNSIG=as.factor(CLNSIG),
  FATHMM=as.factor(FATHMM),
  LRT_pred=as.factor(LRT_pred),
  MutaAss=as.factor(MutaAss),
  MutaTaster=as.factor(MutaTaster),
  PROVEAN=as.factor(PROVEAN),
  Pph2_HDIV=as.factor(Pph2_HDIV),
  Pph2_HVAR=as.factor(Pph2_HVAR),
  SIFT=as.factor(SIFT)
)

dataset_original$MetaSVM <- NULL
dataset_original$CLNSIG <- NULL

# 3) Input data must be encoded for one-hot encoding strategy

to_onehot <- rbind(dataset_original, to_predict)
dummy <- dummyVars(" ~ .", data = to_onehot)
dummy <- data.frame(predict(dummy, newdata = to_onehot))

init = dim(dataset_original)[1] + 1
stop = dim(dummy)[1] + 0
target.to.predict <- dummy[init:stop, ]

#' 4) Load de proposed model

pathmodel <- 'RF13.Rdata'
load(pathmodel)

# 5) Perform prediction with class probabilities
new.predictions <- predict(rf, target.to.predict, type="prob")

# 6) Transform the new predictions in dataframe type.
new.predictions <- as.data.frame(new.predictions)
colnames(new.predictions) <- c("Benign", "Pathogenic")

predict_result = if (new.predictions["Benign"] > new.predictions["
    Pathogenic"]) "B" else "P"
predict_json  = sprintf('{"PREDICT":"%s", "Benign": "%s", "
    Pathogenic":"%s"}'
                        , predict_result, new.predictions["Benign"]
                        , new.predictions["Pathogenic"])

# 7) 'A simple way of printing results.
print(predict_json)
```

Listing 1: R script para uso do AmazonForest

## B.3 Modelos do AmazonForest

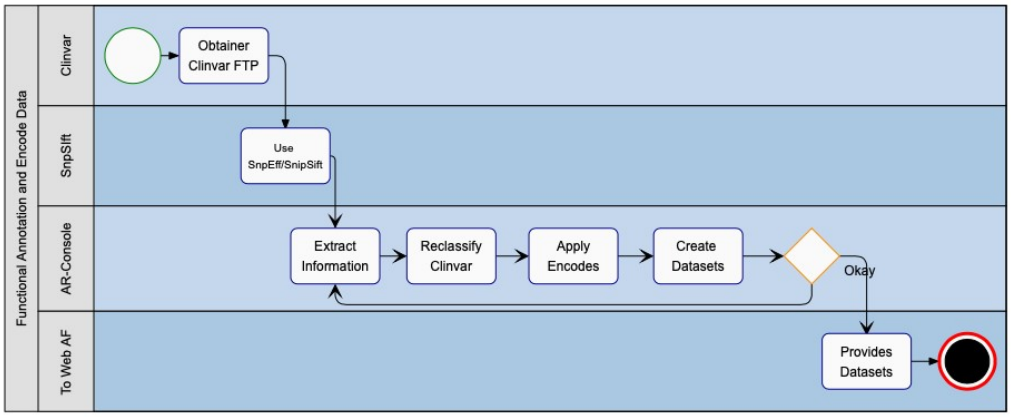



Figura 12: Representação BPMN do fluxo de processo do módulo console do AmazonForest. Desde a sua obtenção dos dados, seu tratamento e dataset final.

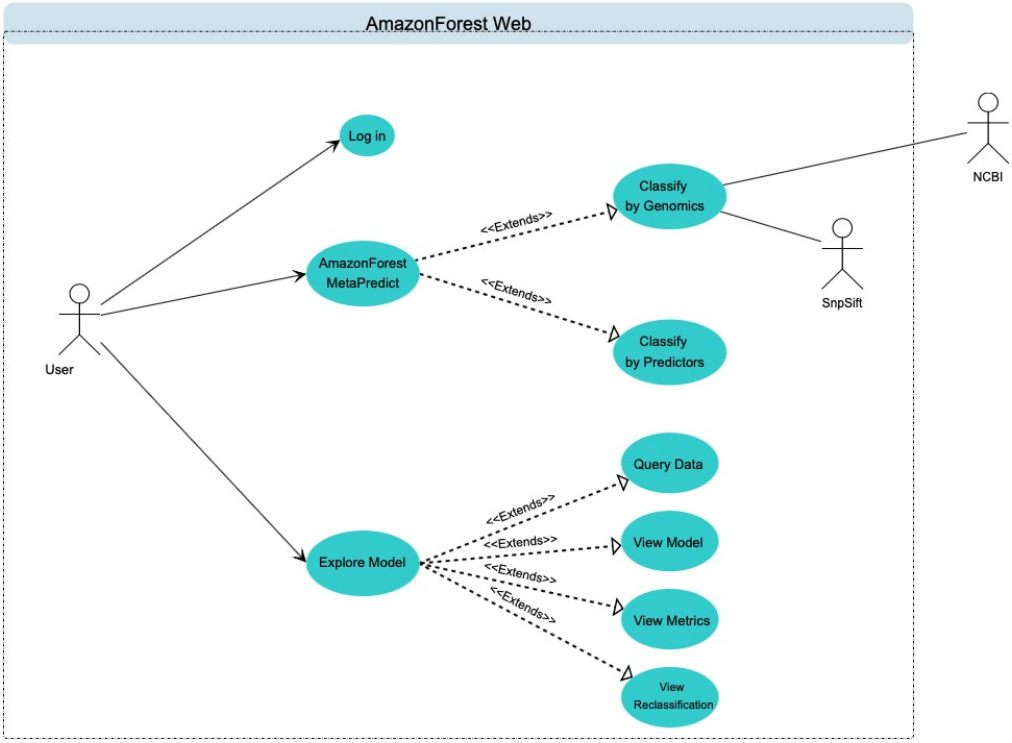



Figura 13: Representação em UML-Caso de Uso do módulo web do AmazonForest. Com suas principais funcionalidades de reclassificação pelo metapreditor bem como a exploração do modelo.

## B.4 Interfaces de usuário do AmazonForest

### B.4.1 Visão Geral

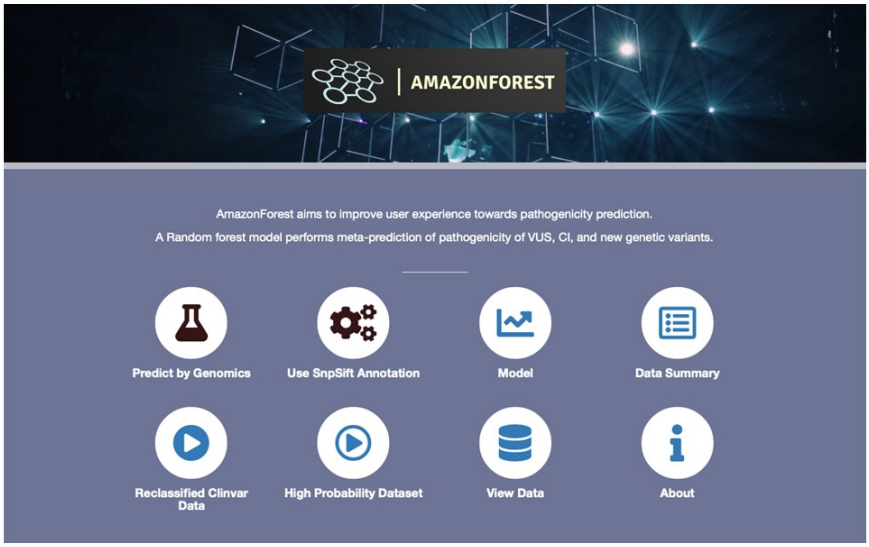



Figura 14: Home AmazonForest.

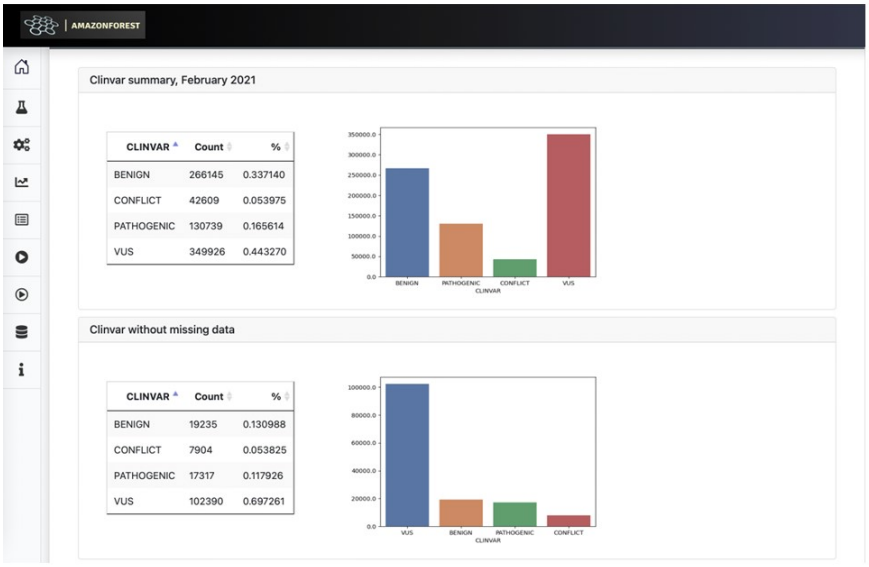
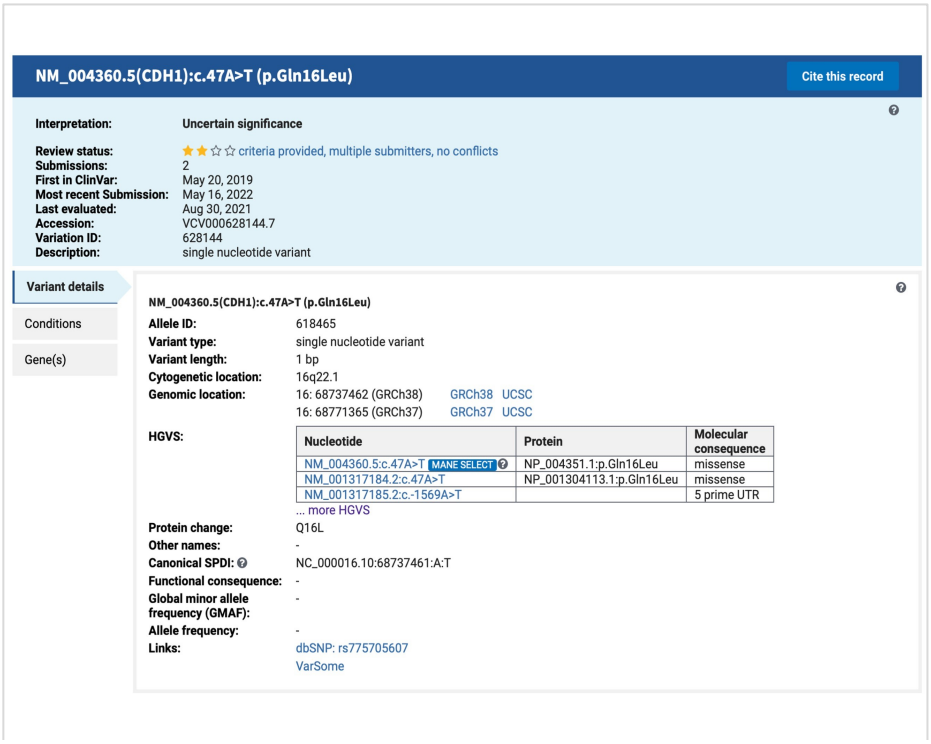


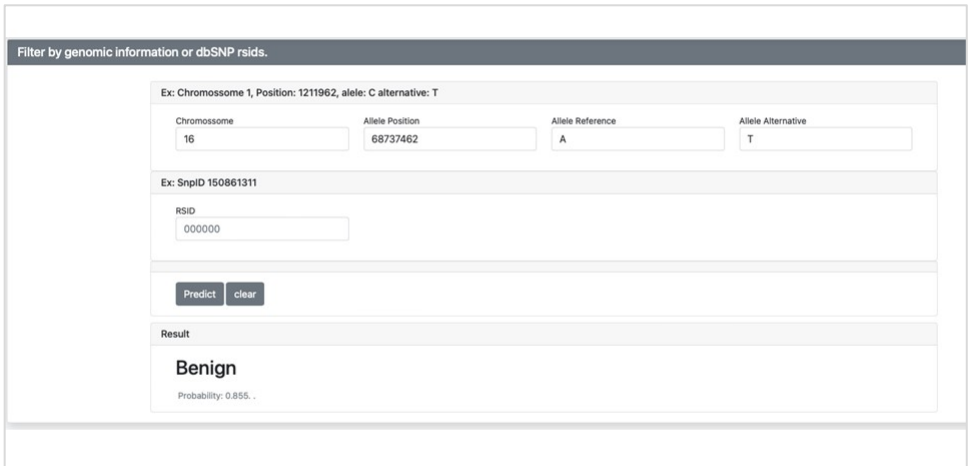

Figura 15: Resumo do conjunto de dados.

B.4.2 Exemplo Consulta

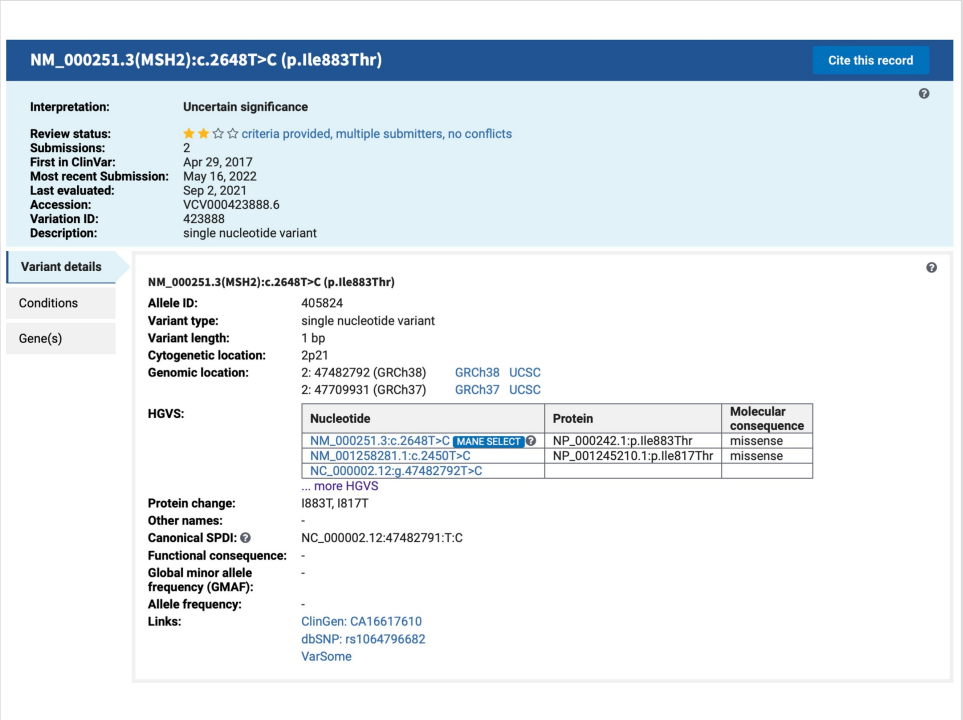
Consulta da variante, chr16:68737462 A>T - NM_004360.5(CDH1):c.47A>T (p.Gln16Leu) no ClinVar.




Consulta da variante por posição genômica no AmazonForest. Chromossome: 16, Allele position: 68737462, Ref: A e Alt: T.



Consulta da variante, chr2:47482792 T>C - NM_000251.3(MSH2):c.2648T>C (p.Ile883Thr) no ClinVar.




Consulta da variante por posição genômica no AmazonForest. Chromossome: 2, Allele position: 47482792, Ref: T e Alt: C.