PREDICCIÓN DE LA PATOGENICIDAD DE LAS VARIANTES MISSENSE DEL GEN TP53

DESARROLLO DE UN METAPREDICTOR PARA LA PATOGENICIDAD DE LAS VARIANTES MISSENSE DEL GEN TP53 MEDIANTE LA METODOLOGIA RANDOM FOREST

Peter Alfonso Olejua Villa, Lic. Mat.

Trabajo de grado presentado a la Facultad de Medicina
como requisito parcial para optar al Grado de
Maestría en Bioestadística
Pontificia Universidad Javeriana
Octubre, 2019

Miembros del Comité de Trabajo de Grado

Carlos Javier Rincón Rodríguez, Estad. MSc.

Estadística y Epidemiología Clínica

Departamento de epidemiología clínica y bioestadística

Pontificia Universidad Javeriana

Fernando Suárez Obando, MD. MS. PhD (c)

Genética Médica e Informática Biomédica

Instituto de Genética Humana

Pontificia Universidad Javeriana

Contenidos

Lista de Tablas

Lista de Figuras

Resumen

Antecedentes: El gen TP53 es un gen supresor de tumores encargado de la apoptosis celular para impedir la formación de células anormales, incluso las cancerosas. Mutaciones genéticas pueden hacer que esta función no se cumpla correctamente, lo que haría que células cancerosas puedan multiplicarse. Frecuentemente mutado en muchos tipos de cáncer, las variaciones genéticas de TP53 más comunes son del tipo missense. Para identificar cuáles variantes pueden ser patogénicas, las guías actuales incluyen el uso de las llamadas herramientas in-silico, modelos predictivos que desempeñan un papel crucial en la priorización de variantes. Las últimas herramientas in-silico recomendadas para las variantes missense de TP53 son los modelos llamados aGVGD, BayesDel y REVEL. Estas dos últimas son herramientas de tipo meta-predictor que no fueron diseñadas específicamente para TP53.

Objetivo: Desarrollar un meta-predictor exclusivamente diseñado para predecir la patogenicidad de las variantes missense del gen TP53.

Materiales y métodos: Se **diseñó** un estudio de desarrollo de herramientas in-silico. Como **población de estudio** se tomaron 1764 variantes missense curadas en la base de datos UMD TP53. De la base de datos dbNSFP se utilizaron 25 herramientas in-silico para el ensamble de un Random Forest, con el que se predijo la patogenicidad de las variantes en cuatro niveles (Benignas, Posiblemente Patogénicas, Probablemente Patogénicas y Patogénicas). Con los datos out-of-bag se estimó el rendimiento del modelo. Mediante un análisis ROC se refinaron los puntos de cortes de la probabilidad de pertenencia a la clase Patogénicas, para utilizar el modelo en dos escenarios distintos de priorización. Los modelos obtenidos fueron aplicados a las variantes missense de significancia incierta (Patogenicidad desconocida) de TP53.

Resultados: Respectivamente, se obtuvo un AUC de 0,99; 0,77; 0.67 y 0.78 para las clases Benignas, Posiblemente Patogénicas, Probablemente Patogénicas y Patogénicas. Del análisis ROC, un primer refinamiento del modelo, para una sensibilidad requerida de 90%, obtuvo un punto de corte de 0.07 para la probabilidad de pertenencia a la clase Patogénicas. En el segundo refinamiento se obtuvo un máximo valor predictivo positivo de 0.67 correspondiente a un punto de corte de 0.66.

Conclusiones: El modelo puede ser utilizado para descartar variantes benignas o para ser utilizado en dos esquemas de priorización distintos. En futuros estudios se puede evaluar el uso del meta-predictor en combinación con otros algoritmos como aGVGD y podría ser recomendado su uso en guías de interpretación de variantes. Las variantes priorizadas como Patogénicas por este modelo pueden ser analizadas en futuros estudios moleculares y poblacionales en investigación en cáncer.

Palabras clave: TP53, Missense, Patogenicidad, Predicción, Random Forest, TP53MiPaPred

Abstract

Background: The TP53 gene is a tumor suppressor gene responsible for cell apoptosis to prevent the formation of abnormal cells, including cancer cells. Genetic mutations can cause this function to not be fulfilled correctly, which would cause cancer cells to multiply. Frequently mutated in many types of cancer, the most common genetic variations of TP53 are of the Missense type. To identify which variants can be pathogenic, current guidelines include the use of so-called in-silico tools, predictive models that play a crucial role in prioritizing variants. The latest in-silico tools recommended for the missense variants of TP53 are the models called aGVGD, BayesDel and REVEL. These last two are meta-predictor tools that were not specifically designed for TP53.

Objective: To develop a meta-predictor exclusively designed to predict the pathogenicity of TP53 missense variants.

Materials and methods: An in-silico tool development study was designed. As a study population, 1764 missense variants cured in the UMD TP53 database were taken. From the dbNSFP database, 25 in-silico tools were used for the assembly of a Random Forest, which predicted the pathogenicity of the variants at four levels (Benign, Possibly Pathogenic, Probably Pathogenic and Pathogenic). With the out-of-bag data, the performance of the model was estimated. Through a ROC analysis, the cut-off points of the probability of membership to the Pathogenic class were refined to use the model in two different prioritization scenarios. The models obtained were applied to the missense variants of uncertain significance (unknown pathogenicity) of TP53.

Results: Respectively, an AUC of 0.99; 0.77; 0.67 and 0.78 was obtained for the Benign, Possibly Pathogenic, Probably Pathogenic and Pathogenic classes. From the ROC analysis, a first refinement of the model, for a required sensitivity of 90%, obtained a cut-off point of 0.07 for the probability of membership to the pathogenic class. In the second refinement a maximum positive predictive value of 0.67 corresponding to a cut-off point of 0.66 was obtained.

Conclusions: The model can be used to rule out benign variants or to be used in two different prioritization schemes. In future studies, the use of the meta-predictor can be evaluated in combination with other algorithms such as aGVGD and its use in variant interpretation guidelines could be recommended. Variants prioritized as Pathogenic by this model can be analyzed in future molecular and population studies in cancer research.

Keywords: TP53, Missense, Pathogenicity, Prediction, Random Forest, TP53MiPaPred

Capítulo 1. Introducción

El gen supresor tumoral TP53, también conocido como "El guardián del genoma", es el gen más estudiado de todos los tiempos (1). Se estima que en más del 50% de los cánceres TP53 está mutado (2). Este gen codifica la proteína p53 que, en condiciones normales y en respuesta a diversas formas de estrés, desencadena actividades como la detención del crecimiento celular, apoptosis o senescencia para evitar la propagación de células aberrantes (3).

Algunas mutaciones genéticas pueden hacer que esta función no se cumpla correctamente, lo que haría que células cancerosas puedan multiplicarse. Las variantes genéticas más comunes en TP53 son missense y somáticas (4). Aunque las mutaciones germinales son poco frecuentes, pueden causar el síndrome de Li-Fraumeni, una predisposición hereditaria a una amplia variedad de tipos de cáncer con inicio temprano (5).

La descripción de las variantes en TP53 está en fase tres (práctica clínica). Se han desarrollado guías clínicas y recomendaciones para el manejo de pacientes en diferentes escenarios; por ejemplo, las variantes se utilizan como biomarcadores en la toma de decisiones para estratificar pacientes para terapias dirigidas (3). En la leucemia linfocítica crónica, algunas mutaciones están asociadas con resistencia a la quimioinmunoterapia y resultados adversos (6). Las variantes también se pueden utilizar como marcadores de pronóstico para mejorar la predicción de sobrevida en ciertos tipos de cáncer (7). La estratificación del riesgo según el estado de TP53, en interacción con otros genes, proporcionará un tratamiento individualizado más eficaz en cáncer de pulmón (8). Los

pacientes con cáncer colorrectal en estadio III y sobreexpresión de la proteína p53 pueden beneficiarse de un tratamiento o seguimiento más agresivo (9).

El tamizaje de variantes en TP53 es ahora común en la práctica clínica y asesoramiento genético. La secuenciación de próxima generación se está utilizando con más frecuencia a medida que disminuye el precio de esta tecnología. Esto ha resultado en el descubrimiento de nuevas variantes de significancia incierta. Estas son variantes cuya patogenicidad aún no ha sido esclarecida y que en la práctica clínica dan lugar a incertidumbre en la toma de decisiones (3,10,11).

Una variante patogénica se define como una que "contribuye mecanicistamente a una enfermedad, pero no es necesariamente totalmente penetrante" (12). Es importante diferenciar este concepto de otros como variantes asociadas, dañinas o deletéreas. Los últimos conceptos se utilizan inconsistentemente como sustitutos de patogenicidad (13).

Esta amplia definición conceptual de una variante patogénica no se refiere a una enfermedad en particular. Por lo tanto, la recomendación ha sido evaluar la patogenicidad de las variantes por enfermedad de interés y por gen (12,14). Se han desarrollado guías para la interpretación de variantes en enfermedades mendelianas (15,16) y para variantes en cáncer (17). Se han realizado adaptaciones de estas guías, específicamente para TP53 (3,6,18), y más están en preparación por paneles de expertos (19).

En estas guías consisten en múltiples criterios que son evaluados sistemáticamente hasta obtener la clasificación patogénica de las variantes. Para el caso particular de TP53, los ensayos funcionales y las herramientas in-silico son criterios de especial importancia (16).

Para las variantes que están parcialmente caracterizadas o de significancia desconocida, no siempre es factible completar la evaluación de cada uno de los criterios para su clasificación. Por lo tanto, en la práctica clínica, las herramientas in-silico juegan un papel crucial el problema general de predecir el impacto funcional de las variantes. Estas son modelos predictivos utilizados para predecir el impacto, el daño, la deleterioridad o la patogenicidad de las variantes. A partir de dicha predicción se realiza la priorización de variantes a estudios moleculares y posteriormente a estudios poblacionales en investigación en cáncer.

Se ha mostrado que estas herramientas pueden no tener concordancia en sus predicciones y que se suelen utilizar de manera inconsistente (20). Por esto, para TP53, se ha recomendado el uso combinado de los algoritmos Align-GVGD, BayesDel y REVEL (21). Sin embargo, de estos, el primero es de uso específico para predecir la actividad transactivacional de las variantes missense de TP53. BayesDel y REVEL son meta-predictores diseñados para predecir sustitutos de patogenicidad en análisis de variantes a gran escala y no incorporan algunos predictores individuales que están disponibles para TP53.

Por tanto, en este estudio, se desarrolla un meta-predictor (ensamble de predictores) para el problema específico de predecir la patogenicidad de las variantes missense del gen TP53. Con inspiración en REVEL, se utilizó la metodología de Random Forest para el ensamble. Para el ajuste se tomaron las variantes missense de la base de datos UMD TP53 cuya patogenicidad está clasificada en cuatro niveles. Este meta-predictor incluye 25 puntuaciones de varias herramientas in-silico contenidas de la base de datos dbNSFP. El nuevo algoritmo recibe por nombre TP53MiPaPred.

Capítulo 2. Marco teórico

2.1 Contexto del problema

El problema general de predecir la patogenicidad de variantes está enmarcado dentro de la medicina de precisión, un enfoque emergente para prevención, diagnóstico y tratamiento de enfermedades que se basa en la variabilidad individual de los pacientes (22–24). El foco de la medicina de precisión es entender la interacción entre las variantes genéticas con factores ambientales y estilos de vida. La Genómica estudia la estructura del genoma y cómo este explica las diferencias entre los humanos (25). Las variantes genéticas pueden influir en la salud. De la misma manera que pueden determinar características físicas como el color de ojos, ellas pueden hacer que individuos sean más o menos susceptibles a enfermedades e igualmente pueden influir en la respuesta a un determinado tratamiento. El objetivo es utilizar la información del genoma de los individuos para la práctica de la medicina y la toma de decisiones (26).

2.2 Priorización

Con los grandes avances en el campo de la secuenciación y la incursión de la medicina en la era del “Big data”, ahora es posible secuenciar el genoma de una persona en poco tiempo y a un bajo costo (relativamente) (27,28). Sin embargo, al aplicar esta tecnología en la práctica clínica resultan una gran cantidad de variantes cuya patogenicidad es de significancia incierta (VUS, por sus siglas en inglés). Para los portadores de estas variantes no existe un manejo

médico establecido. Por lo tanto, determinar la correcta clasificación patogénica de variantes en el entorno clínico es de suma importancia.

Para lograr esta clasificación es necesario realizar estudios de prevalencia poblacional, análisis de segregación, ensayos funcionales y estudios poblacionales que ayuden a evaluar el nivel de patogenicidad de una variante (10). Dada la gran cantidad de variantes VUS, es necesario priorizar variantes que tengan una alta probabilidad de ser patogénicas. Este problema se conoce como priorización de variantes (29).

2.3 Polimorfismos, Mutaciones y Variantes

Estos términos suelen prestarse a bastante confusión en la literatura. Al respecto, se han publicado recomendaciones para su uso (3). En este escrito, los términos, variante y variación son sinónimos y tienen una connotación neutral. El término mutación suele utilizarse para variantes que tienen una consecuencia negativa. Un polimorfismo clásicamente se define como una variación con una frecuencia mayor a 1% en la población.

Se dice que el genoma de un individuo está mutado o contiene una mutación (o variación) cuando su genoma difiere de su correspondiente más frecuente en la población. Una variación es un cambio y el término no debe conllevar una connotación perjudicial. Esto quiere decir que puede haber variantes dañinas, neutrales o incluso beneficiosas. En lo que sigue se utilizará el término variante en lugar de mutación o polimorfismo.

2.4 Variantes Missense

Existen diferentes tipos de variantes según su tamaño: cromosomales, segmentales o puntuales. Esta son un tipo de variante puntual que consiste en la sustitución de un nucleótido y que ocurren en regiones intragénicas de codificación de ADN (exones). Cuando, en el proceso de traducción, la consecuencia es un cambio del anticodón correspondiente la variante se conoce como **missense** (no-sinónima, en algunas traducciones). Se respeta el uso del término missense en inglés dada la preferencia de la comunidad científica por usarlo de esta manera (30).

La sustitución de un nucleótido podría introducir un anticodón de parada y truncar la cadena de polipéptidos. Estas son otro tipo de variantes y se conocen como nonsense (sin-sentido, en algunas traducciones). El foco del presente trabajo son las missense, no las nonsense.(30).

2.5 Patogenicidad

Estos son términos que se prestan a confusión en la literatura (12). Como exponen MacArthur y cols. una variante **patogénica,** conceptualmente, es aquella que "contribuye mecanicistamente a una enfermedad, pero no es totalmente penetrante (puede no ser causa suficiente)". Se resalta que no se refiere a una patología o enfermedad en específico. Operativamente, el nivel de patogenicidad de una variante resulta de la evaluación de un conjunto de criterios, que determinan el nivel de evidencia científica disponible para justificar la priorización de una mutación. Una variante **Implicada** posee cierto nivel de evidencia de ser patogénica.

Mutaciones **Dañinas** son aquellas que por medio de algún mecanismo afectan la función de una proteína. Una variante **Asociada** a alguna enfermedad es significativamente más frecuente, a nivel poblacional, en sujetos con una enfermedad que en aquellos que no la padecen. Una variante **Deletérea** tiende a ser eliminada por selección natural (3).

Por otro lado, para las variantes de TP53 se han realizado ensayos funcionales en levadura para medir la actividad de transactivación (31). La transactivación se suele utilizar como otro sustituto del concepto de patogenicidad o como parte de una definición operativa de una patogenicidad asumida. Sin embargo, transactivación no es patogenicidad (19,32,33). El impacto fenotípico de cada variante debe cuantificarse experimentalmente para estudiar la capacidad supresora de tumores inducida por las variantes. Estudios que evalúan los efectos de las variantes en células humanas se centraron principalmente en variantes que ocurren en puntos críticos (mutational hotspots) de TP53, que representan sólo el 30% de las variantes asociadas al cáncer (34).

2.6 Guías para Clasificación Patogénica

Para evaluar la patogenicidad de las variantes se recomienda el uso de guías, que consisten en un conjunto de criterios para sistematizar el proceso de anotación, clasificación y reporte de las variantes (20).

Por ejemplo, desde el 2015, para enfermedades mendelianas, el consenso para clasificar a una variante como patogénica está plasmado en los **estándares y guías de la ACMG-AMP** (American College of Medical Genetics and the Asociation for Molecular Pathology) y hace énfasis en variantes que ocurren en células germinales (35). Según esta guía, una variante

deberá ser considerada como **Patogénica, Probablemente Patogénica, Significancia desconocida (VUS), Probablemente Benigna o Benigna**. Esto dependerá de la evaluación en conjunto de toda la evidencia disponible para la variante.

Luego de clasificada la variante, se anota su descripción y clasificación en diferentes bases de datos tanto públicas como privadas. Estas bases son curadas y actualizadas periódicamente y pueden ser utilizadas para calibrar o desarrollar nuevas herramientas in-silico. A nivel mundial, los laboratorios siguen estos estándares para el reporte de dichas variantes a médicos genetistas y demás implicados para la toma de decisiones en el ámbito médico.

Las guías de la ACMG-AMP fueron implementadas y posteriormente evaluadas por varios autores, encontrando discordancia entre laboratorios e incluso bases de datos públicas (15,36). Ello causó gran preocupación y controversia alrededor del proceso de anotación y la validez de los reportes resultantes. Un enfoque Bayesiano permitió encontrar que de los 18 criterios planteados en la guías dos son inconsistentes y recomienda futuras mejoras para las guías, pero en general se encontró una consistencia estadísticamente fuerte (37,38).

Particularmente para cáncer, las guías fueron refinadas en Enero del 2017 (17). Es importante hacer notar que este refinamiento se enfocó en variantes somáticas (adquiridas durante la vida y no heredadas) y sus criterios están diseñados para la evaluación de la utilidad clínica y no la patogenicidad de estas variantes. La HGVS (Human Genome Variation Society) también recomienda guías específicas para TP53, que no están enfocadas en la anotación patogénica sino en el impacto funcional de las variantes (3).

Para la evaluación de la patogenicidad, propiamente dicha, dos adaptaciones existen específicamente para TP53. Primeramente, la red para TP53 ERIC (Europen Research

Iniciative on Chronic Lymphocytic Leukemia) recomienda para el estudio de la interpretación de variaciones, la principal base de datos curada exclusivamente para el gen TP53, la base de datos UMD (6). Esta es la base de datos que se usa en este estudio y de donde se obtiene el desenlace a predecir (patogenicidad). La anotación de patogenicidad en esta base de datos se hace mediante otra adaptación de las guías ACMG-AMP. **La clasificación se hace con base en múltiples parámetros predictivos, incluidos predictores de la base de datos dbNSFP, así como la actividad transcripcional basada en el ensayos funcionales basados en levaduras** (31). La adaptación completa aún no se ha descrito completamente y está en preparación según los curadores de la base de datos. A pesar de esto, esta es la única base de datos que se beneficia de la curación experta de datos "artefactuales", lo que la hace ideal para el desarrollo de un nuevo modelo (18). Las variantes son clasificadas como **Patogénica, Probablemente Patogénica, Posiblemente Patogénica, Benigna y VUS (Variant of Unknown Significance).** Para el problema específico, la variable desenlace es una variable ordinal conformada por las primeras cuatro categorías. La "categoría" VUS se entiende como un valor faltante.

Para TP53 existe una segunda **adaptación de las guías de la ACMG-AMP hecha por el grupo ClinGen** (39). Estas especificaciones fueron publicadas en Agosto de 2019 como documento web y aún hace falta un artículo oficial al respecto, como ya existe para el gen PTEN (40). Más adelante se hace referencia a las recomendaciones hechas en estas especificaciones para el uso de in-silico tools.

A manera general, en estas guías las variantes son clasificadas luego de ponderar criterios que se basan en la cantidad de evidencia científica que existe sobre las variantes en estudios poblacionales, ensayos funcionales/computacionales y estudios de segregación familiar. Las

guías presentan dos conjuntos de criterios unos para la clasificación de variantes como benignas y otro para su clasificación como patogénicas. A cada criterio se le da una ponderación desde muy fuerte hasta evidencia de apoyo. Finalmente, las guías exponen unas reglas de cómo se deben combinar los criterios para la clasificación final.

Para el caso particular de TP53, entre otros criterios, las guías incluyen:

- La sustitución de diferentes nucleótidos de ADN conlleva al mismo cambio en el aminoácido.
- Que las mutaciones sean de novo con maternidad y paternidad confirmada.
- Estudios funcionales in vivo o in vitro que cuantifican el efecto dañino de una mutación, como ensayos de transactivación.
- Prevalencia de la variante en individuos afectados versus controles.
- La variante está localizada en el mismo codón que otra variante que es patogénica
- Cosegregación de una enfermedad en múltiples familiares afectados.
- Frecuencia alélica
- Múltiples herramientas in-silico concordantes: aGVGD, Zebrafish, REVEL

2.7 Herramientas In-silico

Los modelos predictivos utilizados para predecir el impacto, el daño, la deleterioridad o la patogenicidad de las variantes y realizar su priorización se conocen como herramientas in-silico (in-silico tools en inglés). Ante la gran cantidad de variaciones que los laboratorios deben analizar y la necesidad de priorizar relativamente pocas mutaciones, los algoritmos predictivos son una herramienta necesaria que aún necesitan mayor desarrollo.

Normalmente, estos se implementan en páginas web y muchas son de libre acceso, lo que hace que sean fáciles de utilizar en la práctica. Existen varias herramientas predictivas y una duda natural es cuál discrimina mejor. Al respecto se han hecho varias comparaciones. Además, también se han integrado o ponderado varios de ellos para formar una única herramienta. Este tipo de algoritmos se conocen como ensamble de herramientas o **meta-predictores**. Aunque en los ajustes originales de los modelos se reporta un buen poder predictivo, en la práctica hay discrepancia entre ellos y se ha encontrado que su poder discriminatorio realmente no es tan alto (13,41–43).

Se han identificado las causas de estos problemas. Primeramente, los datos utilizados para el ajuste de los modelos difieren en la anotación de patogenicidad. Esto se debe al uso inconsistente de otros conceptos sustitutos de patogenicidad que se utilizan para clasificar las variantes con las que se ajustan estos modelos. Además, estos modelos predictivos normalmente son ajustados a nivel genómico, es decir, se busca la aplicación de estos modelos en un gran número de problemas, en donde se estudian varios genes y distintos tipos de variantes. La recomendación ha sido que se ajusten modelos a nivel de genes ya priorizados y de ser posible a nivel de un gen en particular que se esté investigando (44–46). Siguiendo esta recomendación, se delimitó el problema específico a la predicción de la patogenicidad de las variaciones missense del gen TP53.

La mayoría de las herramientas in-silico son ajustadas a un desenlace dicotomizado que, de entrada, no necesariamente refleja el grado de incertidumbre necesario al predecir la patogenicidad de las mutaciones. Por ejemplo, en TP53 las variantes que tienen mayor o menor grado de transactivación son asumidas como benignas o patogénicas, respectivamente; luego se ajustan modelos con esa patogenicidad asumida (19,21). Por esto,

en la presente propuesta se predice directamente la patogenicidad ya anotada en la base de datos UMD TP53 en cuatro niveles de incertidumbre (criterios mencionados arriba).

2.7.1 Tipos de herramientas

Pueden clasificarse según el tipo de desenlace que predicen. La mayoría de las herramientas in-silico predicen dos desenlaces: impacto de la variación en la función proteica (i.e. actividad bioquímica o el control regulatorio de una proteína) o su impacto en el proceso de corte y empalme (splicing) (17).

Algunos algoritmos van más allá e intentan predecir directamente la patogenicidad. Ejemplos de este tipo son modelos modernos, que siguen una metodología de redes neuronales profundas, entre otras alternativas de aprendizaje de máquinas, son REVEL o MVP (47,48). Otros en cambio predicen qué mutaciones asociadas, como DIVAN(49). Otros predicen mutaciones deletéreas como el ya famoso SIFT (50,51).

Es muy fácil, ver distintas definiciones de desenlace confundirse en un mismo modelo, lo que genera desconfianza para los usuarios de estas herramientas (52). Sin embargo, algunos autores ya han comenzado a darle más atención al problema de usar los términos dañino o deletéreo como sustituto de patogenicidad (13).

Según los predictores utilizados en estos modelos, las herramientas in-silico también se pueden dividir en: basados en conservación (efectos de la selección natural negativa), basados en estructura (propiedades fisicoquímicas, localización de la variación), combinados (de los dos anteriores) y meta-predictores (predictores que integran resultados de otros algoritmos) (53). De las comparaciones entre métodos se ha reconocido el poder de los meta-

predictores para la predicción patogénica (13). La presente propuesta corresponde a esta última categoría.

2.7.2 Criterios BP4 y PP3

En las guías de la ACMG-AMP del 2015, en los criterios BP4 y PP3 se menciona que múltiples algoritmos predictivos deberían ser consistentes en su predicción para poder tener evidencia en soporte o no de patogenicidad. Pese a esto, la guía no define qué algoritmos utilizar y cómo combinarlos o ponderarlos. Especial atención se ha reclamado con respecto a este problema (13,20,36). Esto genera discrepancia y confusión entre laboratorios, donde cada uno interpreta estos criterios de manera diferente (54). Este trabajo propone que es en este punto donde un meta-predictor tiene cabida y es particularmente útil (55).

2.7.3 Herramientas in-silico recomendadas para TP53

El panel de expertos del grupo ClinGen, en su adaptación a las guías de la ACMG-AMP para la interpretación de variantes del gen TP53 (39), delimita los criterios BP4 y PP3. La recomendación es usar dos algoritmos Align-GVGD y BayesDel en combinación. Esta recomendación se hizo con base en dos artículos por Fortuno y cols. (19,21). En ellos, para evaluar la capacidad predictiva de los algoritmos se utilizaron datos de referencia en donde la patogenicidad es asumida y dicotomizada a partir de criterios como la transactivación. Para sugerir los mejores puntos de cortes, con dicho conjunto de referencia realizaron análisis de razones de verosimilitud y demostraron que utilizar la combinación de Align-GVGD+BayesDel era superior a la combinación Align-GVGD+REVEL. En su discusión,

los autores mencionan qué mayor cantidad de datos era necesaria para generar razones de verosimilitud confiables. Además, reconocen que los criterios usados para definir los datos de referencia no son universalmente aceptados como criterios estándares para clasificación patogénica de variantes.

Cabe destacar que tanto BayesDel como REVEL, son meta-predictores que no fueron diseñados específicamente para TP53, Por ende, la mejora que aquí se propone lograría un avance natural que iría en línea con el estado del arte del problema específico: Un ensamble específico para las variantes missense de TP53.

2.7.4 Otros algoritmos de predicción In-silico exclusivos de TP53

La documentación de SESHAT, software recomendado por el grupo ERIC, menciona la predicción patogénica mediante algoritmos exclusivamente desarrollados para TP53. Pese a esto, hasta el momento, no es claro si los algoritmos realmente están siendo implementados o están en desarrollo (18,56).

2.7.5 REVEL vs la Patogenicidad anotada en UMD TP53

Una evaluación externa formal del rendimiento de REVEL para predecir la patogenicidad de las variantes missense de TP53 no está dentro de los objetivos de este estudio. REVEL originalmente fue ajustado para predecir un desenlace dicotómico distinto a la patogenicidad ordinal, que es la variable respuesta en este estudio. No obstante, dado que se ha sugerido utilizar este modelo dentro de la clasificación patogénica de estas variantes, un objetivo

secundario de este estudio es comparar descriptivamente el score de REVEL con la patogenicidad ya anotada en UMD TP53.

2.8 Random Forest

Con inspiración en REVEL, en este trabajo se desarrolla un meta-predictor, mediante la metodología de Random Forest (57), para integrar distintas herramientas in-silico y formar una sola predicción. Esta técnica mostró un poder predictivo similar a técnicas más complejas y que están fuera del alcance de ese trabajo como el Boosting o el Aprendizaje Profundo (Deep learning). Mientras que sí ha demostrado ser superior a otras técnicas como regresión, análisis discriminante o máquinas de soporte (13,58,59). Es por esto que se escoge la metodología de Random Forest como técnica de ensamble para el nuevo Meta-predictor que se propone.

El algoritmo de Random Forest (RF) consiste en la agregación de múltiples árboles de decisión. Esta agregación resulta en un mejor rendimiento en predicción comparado con un solo árbol de decisión. Esta técnica puede ser utilizada en problemas de clasificación, regresión o sobrevida. Aquí se expone el RF orientado a la clasificación (60).

El algoritmo original de RF ha tenido variaciones desde su aparición. Por ejemplo, existen implementaciones basadas en árboles de decisión condicionales que podrían ser aplicables a problemas de clasificación con respuestas ordinales (Ordinal Forests) (61,62). Sin embargo, en simulaciones estas variaciones no mostraron una mejora significativa en la capacidad predictiva. Además, el rendimiento fue medido con métricas como el índice Kappa, que no necesariamente aplican a todos los problemas de clasificación. Este estudio se limita al uso

del algoritmo RF original. Se extienden estas ideas en el Capítulo 6Discusión. Se utilizó la implementación de este algoritmo en el paquete "ranger" de R (63).

2.8.1 Algoritmo general

A continuación, se explica en qué consiste el algoritmo de RF de manera general. En los numerales que siguen, se amplía cada paso del algoritmo (64).

1. Se considera un conjunto de datos que conformen una muestra aleatoria de tamaño *n*, llamado datos de entrenamiento (training data): $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$,…, $(\boldsymbol{x}_n, y_n)$. Aquí cada elemento de la muestra *i*, consiste en un vector de *p* predictores $\boldsymbol{x}_i$ y una respuesta categórica $\boldsymbol{y_j}$ cuyas categorías pueden ser $1,\ldots,j,\ldots,J$.
2. Se ajustan un número *ntree* de árboles de probabilidad de la siguiente manera. Para cada árbol:
    2.1. Tomar una muestra aleatoria del conjunto de datos en el punto 1. La muestra puede ser con o sin reemplazo. En muchas implementaciones la muestra suele tomarse con reemplazo de tamaño *n* (Bootstrap sample). Cuando es sin reemplazo, se toma una fracción *sample.fraction* de los n elementos de entrenamiento. El esquema de muestreo se mantiene constante durante toda la construcción del RF (Por ejemplo, para todos los árboles del RF, se hará un muestreo sin reemplazo con una fracción *sample.fraction=0,632*). Este es uno de los hiperparámetros del RF que se optimizan. Suponga que la submuestra tiene tamaño *m.* Las observaciones que no entran en esta submuestra se conocen como los datos *out-of-bag.*

2.2. Se comienza con un nodo raíz que contiene los *m* elementos. Recursivamente se dividen los nodos de manera que cada nodo padre tiene dos hijos (árbol binario). Cada nodo padre del árbol se divide de la siguiente manera:

2.2.1. De los p predictores se escoge aleatoriamente un número *mtry*. Este número se mantiene constante durante toda la construcción del RF (Por ejemplo, para cada árbol RF se escogerán aleatoriamente $mtry = \lceil p \rceil$ predictores). Este es otro de los hiperparámetros del RF.

2.2.2. Para variables continuas, de todas las posibles divisiones con los *mtry* predictores, se escoge la división que maximice el decrecimiento de la impureza de los nodos hijos. La impureza es medida por el índice Gini. En la siguiente sección se expanden estos conceptos. Cuando el índice Gini no decrece usando alguna de las posibles divisiones se deja de dividir el nodo padre actual y se pasa al siguiente nodo que no se haya dividido aún (nodo hermano). Si no hay más nodos por dividir, el árbol está terminado.

2.2.3. Se continúa dividiendo los nodos hasta que cada nodo hoja del árbol (nodos en el último nivel) tenga un tamaño mínimo de *min.node.size*. Esto es otro hiperparámetro a optimizar.

2.8.2 Árboles de Probabilidad

El tipo de árboles utilizados en el algoritmo de RF implementado en este trabajo son árboles de probabilidad (65). Esto son árboles de clasificación no-paramétricos que no tienen ningún supuesto sobre la distribución de los predictores. Son utilizados para calcular la "probabilidad de pertenencia al grupo", en nuestro caso la probabilidad de que una variante pertenezca a

alguno de los niveles de patogenicidad. La forma de calcular las probabilidades por Malley y cols. se ha demostrado que es consistente matemáticamente (64).

Una vez el RF es construido como en el numeral 2.8.1. Para estimar la probabilidad de pertenencia de grupo para una nueva observación *i* se realiza lo siguiente:

1. La observación es pasada por cada uno de los *ntree* árboles de probabilidad.
2. Para cada árbol T_k, en la hoja que resulta la observación, se calculan las proporciones de observaciones que pertenecen a cada clase *j*. Esta es la probabilidad condicional, $P_{Tk}(y_i=j|x_i)$ estimada por el árbol T_k.
3. La probabilidad de pertenencia de grupo estimada por el RF será el promedio de las probabilidades estimadas por cada árbol: $\sum_{k=1}^{ntree} P_{Tk}(y = j|x)/ntree$.

2.8.3 Escogencia de las divisiones, Impureza e índice Gini

Supongamos que el nodo padre N_f tiene n_f observaciones. Las posibles divisiones para un predictor continuo $\boldsymbol{x}$ en este nodo se forman de la siguiente manera. Se ordenan las observaciones por el predictor para obtener los pares ordenados $(x_{(1)},y_{(1)}),\ldots, (x_{(nf)},y_{(nf)})$. Se obtendrán n_f-1 puntos de corte, al tomar los promedios entre los valores $x_{(i)}$ y $x_{(i+1)}$, con i=1: n_{f-1}. La unión de las posibles divisiones de todos los predictores $\boldsymbol{x}_{1,\ldots,}\boldsymbol{x}_p$ conforman el conjunto de todas las posibles divisiones del nodo N_f.

La impureza de un nodo hace referencia al grado de separación que hay en el nodo con respecto a la variable $\boldsymbol{y}$. Por ejemplo, si todas las observaciones del nodo están en la misma categoría y_j, se diría que el nodo es puro; mientras que, si existen observaciones que pertenecen a distintas categorías de la variable *y*, se diría que el nodo es impuro. El grado de

impureza se mide por el índice Gini, mediante la fórmula: $1 - \sum_{j=1}^{J} P_{Nf}^{2}(y = j|x)$. A menor valor del índice Gini mayor pureza. Por ejemplo, cuando todas las observaciones de un nodo pertenecen a una sola categoría y_j el índice toma valor 0.

Suponga que una de las posibles divisiones del nodo N_f tiene dos nodos hijos N_{h1} y N_{h2}, de tamaños n_{h1} y n_{h2}. Sobre cada uno de estos se calcula el índice Gini, G(N_{h1}) y G(N_{h2}). El índice Gini de la división será el promedio ponderado, [n_{h1}* G(N_{h1}) + n_{h1}* G(N_{h1})] / [n_{h1}+n_{h2}].

La mejor división del nodo N_f será aquella que tenga el mínimo índice Gini por debajo de G(N_f). Cuando para ninguna de las posibles divisiones del nodo N_f el Gini decrece, el nodo no se divide más y se pasa al siguiente nodo de la iteración.

2.8.4 Probabilidad de pertenencia de grupo Out-of-bag

En este trabajo no se utilizó ningún conjunto de pruebas (test set), ni validación cruzada debido al bajo tamaño de la muestra en la categoría benigna. En su lugar, las predicciones out-of-bag (OOB) se utilizaron para estimar el error fuera de muestra, i.e. el error que se obtendría en datos no utilizados para entrenar el modelo.

La probabilidad OOB para una variante de entrenamiento se obtiene de igual manera que en 2.8.2, pero utilizando sólo aquellos árboles del bosque que excluyeron la variante de la submuestra con la que se construyó el árbol.

Aunque se ha demostrado que el error OOB puede ser pesimista, en el escenario de este estudio, dado el conjunto de entrenamiento, no es probable que se produzca un sesgo relevante (66). Además, se ha mostrado que en problemas de clasificación, las estimaciones

de OOB pueden ser casi óptimas (60). Estudios futuros podrían utilizar conjuntos de pruebas independientes para la validación y la evaluación comparativa.

2.8.5 Número de árboles *ntree*

El número de árboles no se trató como un hiperparámetro de optimización del RF. Este se estableció tan alto como era computacionalmente factible de acuerdo con las recomendaciones para la optimización de hiperparámetros de RF (60). Para estudiar si el número de árboles es suficiente se estudia la convergencia de alguna métrica a optimizar. En nuestro caso se utilizó la exactitud balanceada (ver numeral 2.8.6).

2.8.6 Exactitud balanceada (balanced accuracy)

Con las probabilidades de pertenencia de grupo es posible predecir la categoría, por ejemplo tomando categoría con máxima probabilidad o mediante alguna dicotomización en especial de estas probabilidades. El error out-of-bag se define como el porcentaje de mala clasificación total de observaciones de entrenamiento. Se ha demostrado que el error OOB puede estar sesgado ante la presencia de clases desbalanceadas (ver numeral 2.9). Como alternativas se han propuesto otras métricas como la exactitud balanceada (BAC, en inglés) (67,68), la perdida logarítmica (logarithmic loss) o el índice Brier (60). En el caso de clases superpuestas (ver numeral 2.9), las últimas dos tienden a sesgarse hacia alguna de las categorías a predecir (69). Con lo cual, será la BAC la métrica para optimizar en la construcción del RF (no para medir su rendimiento, ver el numeral 2.8.9). Esta consiste en el promedio de las sensibilidades de cada clase.

2.8.7 Optimización de Hiperparámetros

Para la optimización de los hiperparámetros del RF, se utilizó el paquete R "tuneRanger". La estrategia implementada en este paquete es una optimización secuencial basada en modelos (SMBO, por sus siglas en inglés) (70). La SMBO iterativamente busca la mejor combinación de hiperparámetros. El algoritmo de optimización implementado consiste en lo siguiente:

1. **Escoger una métrica y una estrategia de evaluación**: En nuestro caso se escogió la BAC para ser evaluada en los datos OOB.
2. **Crear un diseño inicial**, i.e. Escoger una combinación de hiperparámetros aleatoriamente y evaluarlos con la medida y estrategia arriba escogidas. Por defecto se tomaron 30 combinaciones de puntos:
   2.1. *Mtry*: se escoge aleatoriamente de 1,…,p
   2.2. *sample.fraction*: se escoge aleatoriamente de [0.2, 0.9]
   2.3. *min.node.size*: Los valores de tamaño de nodo se muestrean con mayor probabilidad (en el diseño inicial) para valores más pequeños muestreando x desde [0, 1] y transformando el valor mediante la fórmula $[(m*0.2)^{x}]$.
3. Con base en los resultados obtenidos en iteraciones anteriores, **recursivamente**:
   3.1. **Ajustar un modelo de regresión** basada en el diseño de puntos de iteraciones pasadas. La variable dependiente es la métrica de evaluación y las covariables los hiperparámetros.
   3.2. Con el modelo anterior **se propone un nuevo diseño de puntos** que tienen una buena métrica de evaluación esperada. Además, la combinación consiste en puntos que están en regiones del espacio de hiperparámetros en donde no muchos puntos han sido evaluados aún.

3.3. Evaluar la nueva combinación de hiperparámetros y agregarlos a la colección diseño de puntos para posteriores iteraciones.

4. De todas las iteraciones (70 por defecto), se toma el mejor 5% con respecto a la medida de evaluación. Se toma el promedio de los hiperparámetros correspondientes a estas mejores iteraciones.

El esquema de muestreo fue sin reemplazo para evitar sesgos y aumentar ligeramente el rendimiento. Debido a que no se ha demostrado que ninguna regla de división sea superior a otras, se utilizó la división que minimiza la impureza Gini (2.8.3). Se mantuvo la configuración predeterminada con 30 puntos aleatorios evaluados para el diseño inicial y 70 iteraciones en el procedimiento de optimización.

2.8.8 Imputación

RF también puede ser usado para la imputación de datos faltantes en los predictores de entrenamiento. Se ha demostrado que el RF mantiene una buena exactitud incluso cuando el porcentaje de datos faltantes llega a 80% (57). En este estudio todas las variables están por debajo del 4% en datos faltantes. Por otro lado, este método ha demostrado ser robusto ante diferentes mecanismos de aparición de datos faltantes (MCAR, MAR, MNAR) (71).

El método se conoce como método de aproximación. Este consiste en los siguientes pasos:

1. Realizar una imputación inicial. La mediana para predictores continuos.
2. Con la anterior imputación. Refinar la imputación:

   2.1. Ajustar un ramdom forest.

2.2. Construir una matriz de proximidades *n* x *n*: La entrada *ij* cuenta el número de árboles en que las observaciones *i* y *j* están en el mismo nodo hoja.

2.3. Para las observaciones imputadas se actualiza la imputación con el promedio ponderado por los valores de la matriz de proximidades de las observaciones que no fueron imputadas.

2.4. Se repite los pasos 2.1-2.4 hasta que los valores de imputación convergen. Se ha demostrado que típicamente con 10 iteraciones las imputaciones rápidamente convergen. Para analizar la convergencia de las imputaciones es suficiente con ver el error OOB (porcentaje de mala clasificación) en cada iteración. Este se estabiliza cuando las imputaciones convergen.

Para este trabajo la imputación de datos se hace con la implementación del anterior algoritmo hecha en la función *rfImpute* del paquete de R "randomForest".

2.8.9 Importancia de las variables

Aunque no es uno de los objetivos del meta-predictor desarrollado en este trabajo calcular la importancia de las variables como sí lo es en otros problemas de asociación, la metodología de RF permite hacer su cálculo, siendo muy común incluso en problemas de predicción el reporte de los predictores más importantes (47). Sin embargo, se advierte que el proceso de imputación puede afectar la medición de la importancia de las variables en diferentes circunstancias (72). Como el objetivo del presente trabajo es la predicción de una variable respuesta, no se discute sobre el efecto que pueda tener la imputación por el método de proximidad en la importancia de las variables.

La importancia de los predictores en el RF se midió por la importancia de permutación (permutation importance) (57,72). Se ha demostrado que la medida alternativa, importancia Gini es sesgada, tendiendo a inflar la importancia de predictores con muchas categorías o predictores numéricos (70).

Este método considera que un predictor ***x*** es importante si tiene un efecto positivo en el rendimiento predictivo del modelo. Para calcular este efecto positivo se realiza el siguiente procedimiento durante la construcción del RF (72):

1. Para *t = 1,…,ntree*
    a. Sea *i* perteneciente al conjunto O_t de observaciones OOB del árbol *t*. Sea $\hat{y}_{it}$.la predicción del árbol *t* para dicha observación.
    b. Se calcula el porcentaje de observaciones OOB mal clasificadas como $MCR_t \overset{\text{def}}{=} \frac{1}{|O_t|}\sum_{i \in O_t} I(y_{it} \neq \hat{y}_{it})^2$.
    c. Aleatoriamente se permutan los valores del predictor ***x*** para las observaciones OOB y se vuelve a calcular el porcentaje malos clasificados. Sea $MCR_t^{(p)}$ dicho porcentaje.
    d. Calcular del decrecimiento en la exactitud como $D_t = MCR_t^{(p)} - MCR_t$
2. Promediar lo valores obtenidos en 1.d. como $\frac{1}{ntree}\sum_{t=1}^{ntree} D_t$

El valor obtenido es la importancia global en RF de ***x*** por permutación. Por la manera en que se calcula, también se conoce como decrecimiento medio de la exactitud (mean decrease in accuracy, MDA, en inglés).

2.8.10 Medidas de rendimiento

Para el rendimiento, las medidas pertinentes para este problema se basan en la matriz de confusión OOB: sensibilidad, especificidad, valor predictivo positivo (VPP) y valor predictivo negativo. Otras medidas de rendimiento no son de interés en este problema dado que el objetivo es estimar el comportamiento del modelo en escenarios donde se desee priorizar variantes patogénicas y descartar variantes benignas. Por tanto, es importante conocer cada una de esta cuatro medidas y no utilizar métricas que intentan resumir el rendimiento predictivo en un solo escalar (índices Kappa, MCC, F1, BAC, etc.) (68). A continuación, se presentan las fórmulas de las cuatro medidas de rendimiento.

Data un matriz de confusión como la siguiente,

Evento	**Observado**	
Predicho	Positivo	Negativo
Positivo	VP	FP
Negativo	FN	VN

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

$$Especificidad = \frac{VN}{VN + FP}$$

$$Valor\ predictivo\ positivo = \frac{VP}{VP + FP}$$

$$Valor\ predictivo\ negativo = \frac{FN}{VN + FN}$$

Para calcular estas medidas en el problema de clasificación múltiple, se sigue la aproximación “uno vs todos” (one-vs-all, en inglés), i.e. el evento de interés es que una

observación OOB pertenezca o no a una de las clases. De esta manera se calculan las características predictivas para cada una de las clases.

2.8.11 Curvas de características operativas (ROC)

Una vez creado el RF, cada dicotomización de las probabilidades de pertenencia de grupo genera un vector (Probabilidad, Sensibilidad, Especificidad, VPP, VPN). Luego, variando el punto de corte de esta probabilidad, se obtiene una matriz de probabilidades y características operativas. La curva de sensibilidad vs especificidad es una de las más usadas para evaluar la capacidad discriminativa del modelo. Hay una curva por cada probabilidad de pertenencia de grupo. Para cada una de ellas se calcula el área bajo la curva (AUC), que es una medida resumen de que tan bien es distinguida una clase de las demás por el modelo (73). La curva Sensibilidad vs VPP es usada en casos donde el VPP es más importante (74). Ambas curvas son utilizadas en este trabajo como se detalla el numeral 4.8.5.

2.9 Desbalance vs Clases inseparables

El **desbalance** (imbalance, en inglés) entre las clases a predecir se produce cuando una clase tiene muchos menos puntos que otras clases. Las clases se **superponen** (overlap, en inglés) cuando sus muestras tienen características muy similares en términos de los predictores considerados (75). En la literatura de aprendizaje automático, se ha dedicado más atención al problema del desbalance que al problema de la superposición. Sin embargo, se ha demostrado que el desbalance no necesariamente empeora el desempeño de una herramienta de predicción y que la superposición podría ser más importante (76,77). Más adelante vemos

que es esto lo que sucede en este problema caso. Por lo tanto, no se utilizó ninguna técnica de análisis para clases desbalanceadas como aquella basadas en remuestreo, SMOTE o de análisis de costos (78).

Se han propuesto diferentes estrategias para tratar la superposición de clases (79). Es este estudio se optó por la estrategia de "reformular el problema" (80). Se realizó un refinamiento de las predicciones mediante el análisis de las curvas ROC y luego se proponen dos herramientas de predicción diferentes en lugar de proporcionar un modelo único para todos los escenarios. La estrategia empleada se detalla en el numeral 4.8.5.

2.10 Circularidad

Un aspecto importante a mencionar relativo a la evaluación de estas herramientas in-silico es la circularidad, que es un tipo de error en la comparación (benchmarking) de los algoritmos que se ha identificado y ocurre de tres formas (41,81):

1. Tipo 1: Circularidad debido a un solape de entre conjuntos de datos de entrenamiento y validación. Esto aumenta de forma aparente el poder predictivo de ciertos algoritmos, cuando en vez de predecir están recobrando un desenlace con el cual fueron entrenados. Para evitar este tipo de circularidad, al realizar el benchmarking de varios algoritmos sobre un conjunto de datos de validación este deberá ser independiente de todos los conjuntos de datos de entrenamiento de los respectivos algoritmos. Este tipo de circularidad deberá tenerse en cuenta en el futuro cuando se compare la presente propuesta con otras herramientas in-silico.

2. Tipo 2: Ocurre cuando las variantes de un gen en particular son catalogadas predominantemente como dañinas o predominantemente como benignas. Esta es especialmente importante en algoritmos que predicen a nivel genómico. En el presente trabajo, se evita este tipo de circularidad ya que se trabaja en un solo gen, TP53, cuyas variantes han sido curadas previamente.
3. Tipo 3: Ocurre cuando el desenlace, en este caso, patogenicidad de las variantes, ha sido anotada con base en algunos algoritmos in-silico y luego esas variantes son utilizadas en conjuntos de datos para hacer benchmarking. He aquí la importancia de mantener ciertos conjuntos de datos de validación, que además de ser independientes de otros conjuntos de entrenamiento, la patogenicidad haya sido completamente curada.

Cabe destacar que las circularidades están afectando un proceso de comparación en el momento en que se hace el benchmarking entre varios algoritmos. La circularidad no ocurre per se al construir un meta-predictor como algunos autores han propuesto (82).

Capítulo 3. Objetivos

3.1 Propósito del trabajo

Predecir la patogenicidad de las mutaciones missense del gen TP53.

3.2 Objetivo General

Desarrollar un modelo meta-predictor para la anotación patogénica (en cuatro niveles) de variantes missense del gen TP53.

3.3 Objetivos Específicos

- Ajustar un ensamble de herramientas in-silico exclusivo para las variantes missense del gen TP53 para predecir la anotación patogénica (en cuatro niveles), mediante la metodología Random Forest.
- Estimar el rendimiento que tendría el modelo en un escenario práctico.
- Aplicar el modelo seleccionado a las variantes de significancia incierta para identificar variantes que puedan ser priorizadas.
- Implementar el modelo:
    - Distribuir libremente el modelo: código y predicciones por Github.

Capítulo 4. Métodos

4.1 Tipo de estudio

- Investigación básica en genética (83,84).
    - Estudio de variantes genéticas (85).
        - Estudio de consecuencia funcional de variantes genéticas (86).
            - Desarrollo de herramientas in-silico

4.2 Justificación del tipo de estudio

La elección del tipo de estudio es producto de la necesidad de diseñar una herramienta in-silico, específicamente un meta-predictor exclusivo para las mutaciones missense de TP53. Otro tipo de estudios como ensayos moleculares en humanos o GWAS aún no son factibles para todas las variantes debido a la no disponibilidad de recursos. La priorización de estas variantes recae en el uso de herramientas predictivas (desarrollo de herramientas in-silico). La recomendación de la literatura del problema específico es ajustar los modelos para un gen o grupo de genes específicos para que aumente el poder predictivo de estos (13,44–46). Dado que los algoritmos REVEL Y BayesDel son meta-predictores que no fueron creados específicamente para TP53, lo que sigue naturalmente es realizar un estudio de desarrollo de herramientas in-silico en donde se desarrolle un meta-predictor exclusivo para este gen.

4.3 Población

- Unidad de análisis: Variantes genéticas
- Población Blanco: Variantes missense del gen TP53 (somáticas y germinales).
- Población Medible: Variantes missense del gen TP53 curadas en la base de datos UMD TP53 (2017_R2 octubre)

4.4 Criterios de inclusión y exclusión

- Inclusión: Todas las Variantes missense del gen TP53 (tanto somáticas como germinales).
- Exclusión: No aplica

4.5 Tamaño de la muestra

Se toman todas las variantes missense del gen TP53 curadas en la base de datos UMD TP53. Hasta la fecha la base de datos contiene 1764 mutaciones missense únicas de las cuales 1091 son de significancia incierta. Estas últimas no formaran parte del ajuste del modelo, aunque sí de su aplicación. En total se cuenta con 673 mutaciones para modelar.

4.6 Bases de datos

En las guías da la ERIC se recomiendan dos bases de datos para el análisis de variantes TP53 (6,87). Una es la base IARC TP53 (88) y la otra es la base de datos UMD TP53 (89). La primera base fue utilizada previamente para construir datos de referencia de variantes

patogénicas y benignas, cuya patogenicidad es asumida con base en criterios como la transactivación. Del análisis de estos datos surgieron las recomendaciones de usar BayesDel o REVEL (21).

4.6.1 UMD TP53

En este estudio, se utilizó la versión 2017_R2 de la base de datos UMD TP53. Está constituida de 80406 muestras de alteraciones encontradas en tumores, líneas celulares o pacientes con cáncer hereditario en TP53. Contiene variantes germinales y somáticas detectadas en secuenciaciones de próxima generación (next-generation sequences), compilando datos de la literatura y así como de otras bases de datos genómicas más generales. Esta base de datos es la más confiable en la investigación de este gen base, ya que está curada de datos "artefactuales" (89,90). Esta es de libre distribución y está disponible para su descarga.

De estas muestras, 6874 son variantes únicas diferentes, de las cuales, 2063 se clasifican como missense. De estas, 296 son variantes de dinucleótidos. De las 1766 variantes restantes, la variante NM_000546.5:c.993+295C>T (rs1351357911), clasificada en la base de datos como missense, es sinónima. La variante NM_000546.5:c.637C>T (rs1351357911), clasificada en la base de datos como missense, es una variante nonsense. Por lo tanto, hay un total de 1764 variantes únicas de nucleótido único missense en la base de datos. No se aplican criterios de exclusión.

De esta base se utilizaron las variantes con su patogenicidad ya anotada (ver el numeral 2.6). La Tabla 1 en el numeral 4.7.1, muestra la frecuencia de las variantes en la base de datos UMD y su distribución por patogenicidad.

4.6.2 dbNSFP

La lista de herramientas in-silico a ensamblar se obtienen la base de datos **dbNSFP v4.0b1a** (91,92). Esta base también es de libre distribución y es un compilado de distintos puntajes y predicciones a nivel genómico, siendo la referencia para este tipo de estudios.

Las variantes en UMD TP53 fueron cruzadas con los predictores de la base dbNSFP disponibles para variantes missense como sigue. En primer lugar, las variantes se tradujeron de la secuencia de referencia NM_000546.5 a NC_000017.11 (GRCh38) utilizando Mutalyzer 2.0.29 (93). A continuación, se utilizó el algoritmo search_dbNSFP40b1a (disponible con la descarga dbNSFP e implementado en Java) para obtener los predictores de interés. Se seleccionaron un total de 25 puntuaciones. La lista de herramientas in-silico seleccionadas se encuentra en la Tabla 2 del numeral 4.7.34.7.3. La puntuación REVEL también se obtuvo para proporcionar una comparación descriptiva e indirecta con la patogenicidad anotada en UMD TP53. Para predictores con múltiples puntuaciones correspondientes a diferentes transcripciones, se utilizó la transcripción "ENST00000269305", "P04637" como número de adhesión de Uniprot y "P53_HUMAN" como ID de entrada de Uniprot.

4.7 Variables de estudio

4.7.1 Desenlace

El desenlace por predecir es la anotación patogénica, ya curada en UMD TP53 según los estándares y adaptaciones, mencionadas más arriba, de las guías de la ACMG-AMP 2015. La variable es por tanto ordinal con cuatro categorías: Patogénicas > Probablemente

Patogénicas > Posiblemente Patogénicas > Benignas. Con el modelo final se aplicará a las variaciones de significancia incierta (VUS) para su priorización. El tamaño de la muestra por la variable desenlace se distribuye así:

Tabla 1 Distribución de variantes Missense en la base de datos UMD TP53

Patogenicidad	Missense totales en UMD TP3	Missense únicas
Benignas	80 (0.1%)	16 (0.9%)
Posiblemente patogénicas	9,979 (17.2%)	465 (26.4%)
Probablemente patogénicas	5,871 (10.1%)	90 (5.1%)
Patogénicas	38,145 (65.9%)	102 (5.8%)
VUS	3,810 (6.6%)	1,091 (61.8%)
Total	57885	1764

4.7.2 Predictores

La base de datos dbNSFP v4.0b1a contiene una gran cantidad de predicciones para mutaciones a nivel genómico (no solo TP53), conteniendo todos las herramientas in-silico relevantes para el problema específico (94). Para la población en estudio se toman aquellas predicciones que predicen mutaciones dañinas o deletéreas o de efecto splicing. Se tomarán aquellas que no son meta-predictores, i.e. se excluyen los algoritmos que contienen a su vez otros algoritmos como predictores. Esto con fin de evitar la sobrerrepresentación de alguno de los predictores con respecto a los demás dentro del meta-predictor.

Por lo tanto, de los predictores disponibles en dbNSFP v4.0b1a excluimos: MutationTaster (contiene phyloP/phastCons scores) (95), MetaSVM y MetaLR_score (96), M-CAP (97), REVEL(47), MVP (48), MCP (98), DEOGEN2 (contiene PROVEAN) (99), ALoFT

(contiene GERP) (100), CADD (101), DANN (102), EIGEN (103), GENOCANYON (104), Linsight (105).

4.7.3 Definición operativa de las variables a utilizar

A continuación, la definición operativa de las variables que se utilizarán.

Tabla 2 Variables del estudio

Variable	Descripción	Tipo de variable	Rango
chromosome	El número del cromosoma. En este caso constante. El gen TP53 se encuentra en el cromosoma 17. Este valor es utilizado para obtener los predictores de la base dbNSFP	Constante	=17
cDNA_variant	Nomenclatura de mutación según los estándares de HGVS utilizando la secuencia de codificación como referencia (la posición 1 se refiere a la A del ATG de inicio): secuencia de referencia NM_000546.5 Ejemplo: c.782G>A	Identificador	-
WT_AA_1	Aminoácido común: nomenclatura de 1 letra.	Nominal	alanine A arginine R asparagine N aspartic acid D asparagine or aspartic acid B

			cysteine C glutamic acid E glutamine Q glutamine or glutamic acid Z glycine G histidine H isoleucine I leucine L lysine K methionine M phenylalanine F proline P serine S threonine T tryptophan W tyrosine Y valine V
Mutant_AA_1	Aminoácido mutante: nomenclatura de 1 letra	Nominal	alanine A arginine R asparagine N aspartic acid D asparagine or aspartic acid B cysteine C glutamic acid E glutamine Q glutamine or glutamic acid Z glycine G histidine H isoleucine I leucine L lysine K methionine M phenylalanine F proline P serine S threonine T

			tryptophan W tyrosine Y valine V
position	Transformada de la posición de la mutación de NM_000546.5 a NC_000017.11 (GRCH38 o HG38). Para la transformación se utiliza el conversor de posiciones Mutalyzer 2.0.29 (93,106). Esta será la posición que se utilizará para cruzar con la base de datos de dbNSFP	Identificador	7669614 a 7676594
ref	Base nitrogenada de referencia en NC_000017.11. Esta será la posición que se utilizará para cruzar con la base de datos de dbNSFP.	Nominal	Adenina A Guanina G Timina T Citosina C
alt	Base nitrogenada variante en NC_000017.11. Esta será la posición que se utilizará para cruzar con la base de datos de dbNSFP	Nominal	Adenina A Guanina G Timina T Citosina C
Ensembl_transcriptid	ID de transcripción de Ensembl (entradas múltiples separadas por ";"). Se utiliza para obtener algunos scores de dbNSFP (107).	Constante	= ENST00000269305
Ensembl_proteinid	IDs de proteínas Ensembl Múltiples entradas separadas por ";",	Constante	= ENSP00000269305

	correspondientes a Ensembl_transcripti d. Se utiliza para obtener algunos scores de dbNSFP (107).		
Uniprot_acc	Número de acceso de Uniprot que coincide con el Ensembl_proteinid Múltiples entradas separadas por ";". Se utiliza para obtener algunos scores de dbNSFP (107).	Constante	= P04637
Uniprot_entry	ID de entrada de Uniprot que coincide con Ensembl_proteinid Múltiples entradas separadas por ";". Se utiliza para obtener algunos scores de dbNSFP (107).	Constante	= P53_HUMAN
SIFT_score	Efecto funcional predicho utilizando algoritmo SIFT (108) SIFT calcula la probabilidad de que un aminoácido en una posición sea tolerado condicionalmente al aminoácido más frecuente que se tolera. Está basado en homología de secuencias y propiedades físicas de los aminoácidos. Mientras más pequeña la probabilidad mayor probabilidad de daño. Si es menor a	Cuantitativ a continua	0 a 1

	0.05 se considera dañino, si no se considera tolerable. Múltiples puntuaciones separadas por ";" correspondientes a Ensembl_proteinid.		
Polyphen2_HDIV_score	PolyPhen-2 calcula la probabilidad posterior de Naïve Bayes de que una mutación sea Dañina (109). Fue ajustado sobre dos bases de datos. Este primer score está ajustado en la base de datos HumDiv. Múltiples entradas separadas por ";", correspondientes a Uniprot_acc.	Cuantitativa continua	0 a 1
Polyphen2_HVAR_score	Polyphen2 basada en HumVar (109). Múltiples entradas separadas por ";", correspondientes a Uniprot_acc.	Cuantitativa continua	0 a 1
LRT_score	El p-valor de una prueba de razón de verosimilitudes de dos colas para la significancia de la conservación de posiciones de aminoácidos dentro del proteoma humano (110).	Cuantitativa continua	0 a 1
MutationAssessor_score	Predice el impacto funcional de las sustituciones de aminoácidos en proteínas basadas en la conservación evolutiva del	Cuantitativa continua	-5.17 a 6.49

	aminoácido afectado en homólogos de proteínas. El score de impacto funcional (FIS) es una combinación de un score de conservación y uno de especificidad (111). Las entradas múltiples están separadas por ";", correspondientes a Uniprot_entry.		
FATHMM_score	Predice los efectos funcionales de las mutaciones de la proteína missense combinando la conservación de secuencias dentro de los modelos ocultos de Markov (HMM), que representan la alineación de secuencias homólogas y dominios proteicos conservados, con "pesos de patogenicidad", que representan la tolerancia general de la proteína / dominio a las mutaciones. A menor score más probabilidad de que la variación sea dañina (112). Múltiples puntuaciones separadas por ";", correspondientes a Ensembl_proteinid.	Cuantitativa continua	-16.13 a 10.64

fathmm-MKL_coding_score	Considerado una mejora al anterior, incluyendo variaciones en regiones de codificación y no-codificación, es un p-valor para predecir la deleterioridad o neutralidad de las variaciones. Mientras más cercano a 1 mayor confianza de deleterioridad (113).	Cuantitativa continua	0 a 1
fathmm-XF_coding_score	Considerado una mejora del MKL al incluir más predictores. Es un p-valor para predecir la deleterioridad o neutralidad de las variaciones. Mientras más cercano a 1 mayor confianza de deleterioridad (114).	Cuantitativa continua	0 a 1
PROVEAN_score	PROVEAN predice si una sustitución de aminoácidos o indel tiene un impacto en la función biológica de una proteína. El Delta alignment score, basado en alineación, mide el cambio en la similitud de secuencia a un homólogo de secuencia de proteínas antes y después de la introducción de una	Cuantitativa continua	-14 a 14

	variación de aminoácidos. A menor score mayor efecto “deletéreo”. Un mayor score sugiere un efecto neutro (115). Múltiples puntuaciones separadas por ";", correspondientes a Ensembl_proteinid.		
VEST4_score	es un Random Forest que agregando p-valores predice la importancia funcional de las mutaciones missense en función de la probabilidad de que sean patogénicas (116,117). Cuanto mayor sea la puntuación, más probable es que la mutación pueda causar un cambio funcional. Múltiples puntuaciones separadas por ";", correspondientes a Ensembl_transcripti d.	Cuantitativa continua	0 a 1
MutPred_score	El modelo MutPred2 es un modelo basado en secuencias que consiste en un ensamble de 30 redes neuronales. El modelo arroja dos scores, de los cuales	Cuantitativa continua	0 a 1

	se considera el "general score", el cual indica la patogenicidad de la variante (118).		
PrimateAI_score	Una puntuación de predicción de patogenicidad para variantes missense basadas en variantes comunes de especies de primates no humanos que utilizan una red neuronal profunda (82).	Cuantitativa continua	0 a 1
integrated_fitCons_score	Un score de adecuación biológica (fitness) que a mayor puntaje indica que una mayor proporción de sitios nucleicos de la clase funcional a la que pertenece la posición genómica están bajo presión selectiva, por lo que es más probable que tengan una importancia funcional (119). Este score es una integración de los siguientes tres:	Cuantitativa continua	0 a 1
GM12878_fitCons_score	El Score de fitCons basado en células linfoblastoides (119).	Cuantitativa continua	0 a 1
H1-hESC_fitCons_score	El Score de fitCons basado en células humanas madre embrionarias H1 (119).	Cuantitativa continua	0 a 1
HUVEC_fitCons_score	El Score de fitCons basado en células	Cuantitativa continua	0 a 1

	epiteliales de la vena umbilical humana (119).		
GERP++_RS	GERP ++ RS es un score de conservación, cuanto mayor sea la puntuación, más conservado será el sitio de la variación (120).	Cuantitativa continua	-12.3 a 6.17
phyloP100way_vertebrate	El score de conservación phyloP se basa en las lineaciones múltiples de 100 genomas de vertebrados (incluido el humano). Cuanto mayor sea la puntuación, más conservado es el sitio de la variación (121)	Cuantitativa continua	-20.0 a 10.003
phyloP30way_mammalian	El score de conservación phyloP se basa en las lineaciones múltiples de 30 genomas de mamíferos (incluido el humano). Cuanto mayor sea la puntuación, más conservado es el sitio de la variación (121)	Cuantitativa continua	-20 a 1.312
phyloP17way_primate	El score de conservación phyloP se basa en las lineaciones múltiples de 17 genomas de primates. Cuanto mayor sea la	Cuantitativa continua	-13.362 a 0.756

	puntuación, más conservado es el sitio de la variación (121)		
phastCons100way_vertebrate	El score de conservación phastCons se basa en las lineaciones múltiples de 100 genomas de vertebrados (incluido el humano). Cuanto mayor sea la puntuación, más conservado es el sitio de la variación (122)	Cuantitativa continua	0 a 1
phastCons30way_mammalian	El score de conservación phastCons se basa en las lineaciones múltiples de 30 genomas de mamíferos (incluido el humano). Cuanto mayor sea la puntuación, más conservado es el sitio de la variación (122)	Cuantitativa continua	0 a 1
phastCons17way_primate	El score de conservación phastCons se basa en las lineaciones múltiples de 17 genomas de primates. Cuanto mayor sea la puntuación, más conservado es el sitio de la variación (122)	Cuantitativa continua	0 a 1
SiPhy_29way_logOdds	Score de conservación basado en 29 genomas de	Cuantitativa continua	0 a 37.9718

	mamíferos. A mayor score mayor conservación de la posición.		
bStatistic	Estimación del valor de selección de fondo basado en conservación de secuencias en mamíferos (123). Los valores cercanos a 0 representan la eliminación casi completa de la diversidad como resultado de la selección de fondo y los valores cercanos a 1000 que indican la ausencia de selección de fondo.	Cuantitativa continua	0 a 1000
Patogenicidad (desenlace)	Anotación patogénica ya curada en UMD TP53 de acuerdo con las adaptaciones hechas a las guías ACMG-AMP 2015 y ERIC disponibles en la documentación de Seshat más arriba referenciada.	Ordinal	1. Benignas 2. Posiblemente patogénicas 3. Probablemente patogénicas 4. Patogénicas Las variantes VUS no tomarán para al ajuste del modelo. El modelo será aplicado a estas para priorizarlas.

4.8 Análisis estadístico

4.8.1 Análisis descriptivo

Para la descripción de los predictores, que son variables continuas, se utilizó como medidas de resumen de tendencia central y dispersión, la mediana y el rango intercuartil

respectivamente. Para las variables categóricas y para el conteo de datos faltantes, se utilizó frecuencia absolutas y relativas, i.e. conteos y porcentajes respectivamente.

4.8.2 Descripción del comportamiento de REVEL en UMD TP53

El puntaje de REVEL también fue obtenido de la base dbNSFP, para realizar una comparación descriptiva contra el desenlace patogenicidad. Se tabuló la dicotomización utilizando el corte sugerido de 0.5 (21), contra la patogenicidad anotada en la base de datos UMD TP53. También se realizó un diagrama de cajas del puntaje contra las categorías de patogenicidad. No se reportó ninguna medida de rendimiento predictivo para REVEL para evitar confusiones con un benchmarking formal.

4.8.3 Ajuste del Random Forest

Para el ensamble de herramientas in-silico se utilizó la metodología de Random Forest (RF). El ajuste consta de las siguientes fases:

1. Imputación de los datos: Se utilizó el método de aproximación iterando el algoritmo 20 veces y en donde los RF constaban de 10001 árboles. La convergencia se analiza mediante los errores OOB de cada iteración.
2. Optimización de los parámetros: Se utilizó la metodología SMBO implementada en el paquete tuneranger de R.
3. Se ajusta el RF final con los parámetros obtenidos.

4.8.4 Estimación del error fuera de muestra.

Para estimar el comportamiento del RF en un escenario práctico, el error fuera de muestra se estimó utilizan las probabilidades OOB, la clase de patogenicidad predicha por el RF es aquella cuya probabilidad de pertenencia de grupo es mayor. Con las clases predichas y las clases observadas, se construyó una matriz de confusión con la que se valuó las medidas de rendimiento (sensibilidad, especificidad, y valores predictivos).

4.8.5 Refinamiento del modelo

Dado que las clases patogénicas son inseparables, se utiliza la estrategia de “reformular el problema” (80) y se proponen dos escenarios con intereses distintos. Uno donde el objetivo es priorizar pocas variantes que tengan una alta probabilidad de ser patogénicas; y otro escenario donde el interés es no perder ninguna variante patogénica en la priorización. En el primero el interés es tener un valor predictivo positivo alto y en el segundo tener una sensibilidad alta.

Mediante un análisis de las curvas de características operativas, se realiza un refinamiento post-hoc para encontrar el mejor punto de corte para clase patogénica y se proponen dos modelos finales para cada uno de los escenarios. Los datos OOB se utilizaron como un "conjunto de validación" para estimar el error de predicción y para la selección de los puntos de corte (124).

Para el modelo con mayor sensibilidad en la clase patogénica, se seleccionó la probabilidad de pertenencia de grupo correspondiente a una sensibilidad requerida del 90%. Las variantes con una probabilidad de pertenencia a la clase patogénicas superior a ese punto corte son

predichas como patogénicas. Dado que este refinamiento también afecta a la predicción en otras clases, la predicción predeterminada por el RF se utiliza para la clase benigna. Luego de tener estas dos categorías predichas, para predecir las otras dos clases, se escoge la clase con mayor probabilidad de pertenencia de grupo predeterminada por RF. Este modelo de predicción se llamó "TP53MiPaPred_sens".

Para el modelo con mayor valor predictivo positivo, el corte se seleccionó inspeccionando la curva de valor predictivo positivo contra sensibilidad para la clase patogénica. Se seleccionó el punto de corte correspondiente al VPP más alto por debajo de uno (máximo local por debajo de uno). Este modelo de predicción se llamó "TP53MiPaPred_ppv".

4.8.6 Aplicación de los modelos a las VUS

Los modelos TP53MiPaPred_ppv y TP53MiPaPred_sens fueron utilizados para predecir la patogenicidad de las variantes VUS y sus resultados fueron comparados entre sí y con la dicotomización de REVEL. Para la comparación se utilizan tablas de contingencia 2x2 y frecuencias absolutas y relativas. Se tomaron solo las variantes VUS con predictores completos. Si bien es posible hacer la imputación por RF o cualquier otro método. Los mecanismos de la aparición de datos faltantes deben ser abordados en futuros estudios.

Como software estadístico se utilizará R versión 3.5.1 (125).

Capítulo 5. Resultados

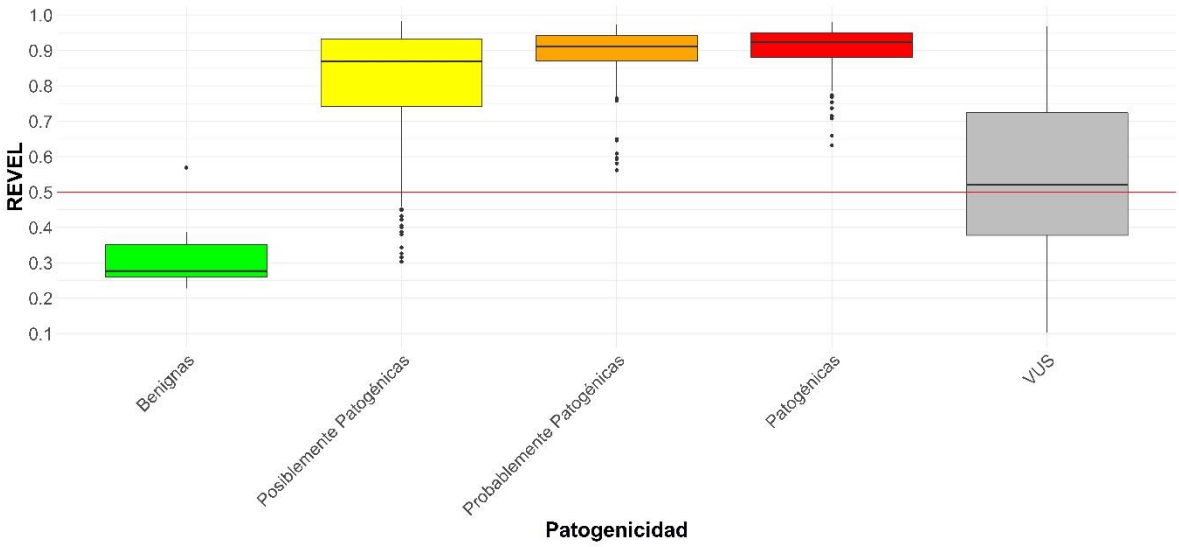
5.1 Descripción del comportamiento de REVEL en UMD TP53

El punto de corte sugerido para REVEL es de 0.5. Al aplicar este punto de corte a las variantes missense de TP53 los resultados se presentan en la Tabla 3. Esencialmente al aplicar esta dicotomización REVEL es capaz de distinguir las variantes benignas. Aproximadamente 4% de las variantes posiblemente patogénicas serían clasificadas erróneamente como benignas. Sin embargo, al aplicar la dicotomización a las variantes de significancia incierta, aproximadamente el 47% de estas serían clasificadas como Benignas. Un número muy alto considerando la distribución de las variantes por clase patogenicidad en la Tabla 1. En la Figura 1 se aprecia que el puntaje no logra separar las otras tres clases patogenicidad.

Tabla 3 REVEL dicotomizado vs patogenicidad anotada en UMD TP53

	Total	**[0,0.5]**	**(0.5,1]**
Patogenicidad	**No. 1,764**	**No. 542**	**No. 1,222**
Benignas	16 (0.9%)	15 (93.8%)	1 (6.2%)
Posiblemente patogénicas	465 (26.4%)	18 (3.9%)	447 (96.1%)
Probablemente patogénicas	90 (5.1%)	0 (0.0%)	90 (100.0%)
Patogénicas	102 (5.8%)	0 (0.0%)	102 (100.0%)
VUS	1,091 (61.8%)	509 (46.7%)	582 (53.3%)

Figura 1 REVEL vs patogenicidad anotada en UMD TP53



5.2 Resumen de los predictores a ensamblar

En la Tabla 4 se presenta el resumen de los predictores por clase de patogenicidad y en total. Varios de los predictores difieren en la media para clase benigna, siendo esta la más diferenciable. Existe un bloque de 13 variantes que tienen datos faltantes en los 8 primeros predictores. Sin embargo, esta clase no tiene datos faltantes en el resto de los 17 predictores. En términos de predicción, en lo que sigue más abajo, esto no representó un problema. En la

Tabla 5 se aprecia que 10 de 25 predictores tienen datos faltantes. La variable con más datos faltantes es *MutPred_score*; sin embargo, su porcentaje de datos faltantes está por debajo del 4%. En la Tabla 6 se muestra que, excepto la clase benignas, el resto tiene un porcentaje de casos completos superior a 95%.

Tabla 4 Resumen de los predictores

Predictor	Benignas	Posiblemente patogénicas	Probablemente patogénicas	Patogénicas	VUS	Total
	No. 16	**No. 465**	**No. 90**	**No. 102**	**No. 1,091**	**No. 1,764**
SIFT						
Mediana (RIQ)	0.26 (0.21,0.45)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.06 (0.00,0.24)	0.01 (0.00,0.12)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	13 (0.74%)
FATHMM						
Mediana (RIQ)	-5.45 (-5.48, -5.44)	-6.97 (-7.27, -6.72)	-7.29 (-7.41, -6.99)	-7.30 (-7.50, -7.04)	-6.31 (-6.73, -5.49)	-6.69 (-7.05, -5.56)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	13 (0.74%)
PROVEAN						
Mediana (RIQ)	-0.14 (-0.18,0.15)	-4.88 (-6.55, -2.89)	-6.21 (-7.75, -3.89)	-6.17 (-8.43, -3.94)	-1.17 (-3.03, -0.19)	-2.62 (-5.11, -0.66)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	13 (0.74%)
VEST4						
Mediana (RIQ)	0.08 (0.08,0.12)	0.82 (0.65,0.90)	0.92 (0.81,0.96)	0.93 (0.87,0.96)	0.36 (0.22,0.57)	0.56 (0.30,0.82)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	13 (0.74%)
Polyphen2 HDIV						
Mediana (RIQ)	0.01 (0.00,0.05)	1.00 (0.95,1.00)	1.00 (0.99,1.00)	1.00 (0.99,1.00)	0.16 (0.00,0.96)	0.87 (0.02,1.00)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	13 (0.74%)
Polyphen2 HVAR						
Mediana (RIQ)	0.00 (0.00,0.02)	0.99 (0.84,1.00)	1.00 (0.96,1.00)	0.99 (0.98,1.00)	0.19 (0.01,0.88)	0.77 (0.06,0.99)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	13 (0.74%)
MutationAssessor						
Mediana (RIQ)	0.97 (0.96,1.16)	2.88 (2.35,3.18)	3.16 (2.90,3.27)	3.20 (2.93,3.29)	1.93 (1.32,2.51)	2.36 (1.67,2.98)
Faltante	13 (81.25%)	0 (0%)	1 (1.11%)	0 (0%)	3 (0.27%)	17 (0.96%)
LRT						
Mediana (RIQ)	0.36 (0.18,0.37)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.00)	0.00 (0.00,0.13)	0.00 (0.00,0.03)
Faltante	13 (81.25%)	0 (0%)	0 (0%)	0 (0%)	10 (0.92%)	23 (1.30%)
fathmm MKL coding						
Mediana (RIQ)	0.05 (0.03,0.43)	0.99 (0.96,0.99)	0.99 (0.98,1.00)	0.99 (0.98,0.99)	0.79 (0.18,0.97)	0.95 (0.40,0.99)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
fathmm XF coding						
Mediana (RIQ)	0.06 (0.05,0.09)	0.90 (0.64,0.95)	0.93 (0.81,0.96)	0.94 (0.85,0.96)	0.31 (0.15,0.68)	0.57 (0.21,0.91)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
MutPred						
Mediana (RIQ)	0.28 (0.20,0.55)	0.85 (0.73,0.92)	0.91 (0.86,0.96)	0.94 (0.89,0.97)	0.47 (0.26,0.70)	0.68 (0.38,0.86)
Faltante	4 (25.00%)	16 (3.44%)	1 (1.11%)	5 (4.90%)	24 (2.20%)	50 (2.83%)
PrimateAI						
Mediana (RIQ)	0.34 (0.30,0.37)	0.60 (0.49,0.69)	0.66 (0.60,0.72)	0.69 (0.59,0.74)	0.44 (0.35,0.54)	0.50 (0.39,0.63)
Faltante	14 (87.50%)	4 (0.86%)	1 (1.11%)	0 (0%)	58 (5.32%)	77 (4.37%)
integrated fitCons						
Mediana (RIQ)	0.67 (0.64,0.67)	0.72 (0.72,0.72)	0.72 (0.72,0.72)	0.72 (0.72,0.72)	0.72 (0.72,0.72)	0.72 (0.72,0.72)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
GM12878 fitCons						
Mediana (RIQ)	0.70 (0.67,0.70)	0.70 (0.70,0.70)	0.70 (0.70,0.70)	0.70 (0.70,0.70)	0.70 (0.70,0.70)	0.70 (0.70,0.70)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
H1.hESC fitCons						
Mediana (RIQ)	0.61 (0.61,0.61)	0.70 (0.70,0.72)	0.70 (0.70,0.72)	0.72 (0.70,0.72)	0.70 (0.70,0.72)	0.70 (0.70,0.72)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
HUVEC fitCons						
Mediana (RIQ)	0.66 (0.32,0.69)	0.74 (0.74,0.74)	0.74 (0.74,0.74)	0.74 (0.74,0.74)	0.74 (0.74,0.74)	0.74 (0.74,0.74)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
GERP++ RS						
Mediana (RIQ)	0.77 (-0.64,1.21)	4.62 (3.53,5.13)	4.62 (3.64,5.28)	4.62 (3.64,5.13)	2.92 (0.29,4.45)	3.65 (1.40,4.71)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
phyloP100way vertebrate						
Mediana (RIQ)	-0.04 (-0.33,0.46)	6.94 (2.62,8.18)	7.91 (5.31,9.30)	7.91 (5.30,7.91)	1.14 (0.13,3.79)	2.51 (0.44,7.47)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
phyloP30way mammalian						
Mediana (RIQ)	0.13 (-0.41,0.73)	1.14 (1.02,1.30)	1.14 (1.02,1.18)	1.03 (1.02,1.14)	1.00 (0.11,1.14)	1.03 (0.18,1.17)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
phyloP17way primate						
Mediana (RIQ)	0.47 (-0.20,0.67)	0.67 (0.60,0.75)	0.67 (0.60,0.68)	0.60 (0.60,0.66)	0.66 (0.35,0.67)	0.66 (0.60,0.68)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

phastCons100way vertebrate						
Mediana (RIQ)	0.00 (0.00,0.02)	1.00 (1.00,1.00)	1.00 (1.00,1.00)	1.00 (1.00,1.00)	0.68 (0.00,1.00)	1.00 (0.02,1.00)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
phastCons30way mammalian						
Mediana (RIQ)	0.22 (0.06,0.66)	0.97 (0.83,1.00)	0.98 (0.84,1.00)	0.98 (0.87,1.00)	0.72 (0.03,0.98)	0.90 (0.10,0.99)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
phastCons17way primate						
Mediana (RIQ)	0.97 (0.86,0.99)	0.97 (0.74,1.00)	0.98 (0.75,1.00)	0.98 (0.84,1.00)	0.77 (0.21,0.99)	0.91 (0.55,0.99)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
X29way logOdds						
Mediana (RIQ)	4.24 (3.38,4.87)	12.31 (9.50,13.83)	12.94 (10.68,15.36)	13.59 (10.88,15.66)	8.18 (5.20,11.98)	9.98 (6.46,13.06)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	7 (0.64%)	7 (0.40%)
bStatistic						
Mediana (RIQ)	439.00 (436.00,439.00)	433.00 (433.00,434.00)	433.00 (433.00,434.00)	433.00 (432.00,434.00)	434.00 (433.00,434.00)	434.00 (433.00,434.00)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
REVEL						
Mediana (RIQ)	0.28 (0.26,0.35)	0.87 (0.74,0.93)	0.91 (0.87,0.94)	0.92 (0.88,0.95)	0.52 (0.38,0.72)	0.69 (0.46,0.88)
Faltante	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Tabla 5 Datos faltantes por variables en los datos de entrenamiento

Variable	**Faltantes**	**%**
MutPred_score	26	3.9
PrimateAI_score	19	2.8
MutationAssessor_score	14	2.1
SIFT_score	13	1.9
FATHMM_score	13	1.9
PROVEAN_score	13	1.9
VEST4_score	13	1.9
Polyphen2_HDIV_score	13	1.9
Polyphen2_HVAR_score	13	1.9
LRT_score	13	1.9

Tabla 6 Casos completos por clase de patogenicidad en los datos de entrenamiento

Patogenicidad	**Total**	**Casos completos**
	No. 673	**No. 632**
Benignas	16 (2.4%)	0 (0.0%)
Posiblemente patogénicas	465 (69.1%)	447 (96.1%)
Probablemente patogénicas	90 (13.4%)	88 (97.8%)
Patogénicas	102 (15.2%)	97 (95.1%)

5.3 Imputación

Luego de correr el método de proximidades, vemos los errores OOB por cada iteración y cada clase de patogenicidad en la Tabla 7. Desde las primeras iteraciones el OOB se mantuvo alrededor de 28.23%

Tabla 7 Convergencia del método de proximidades para imputación

Iteración	OOB	Benignas	Posiblemente patogénicas	Probablemente patogénicas	Patogénicas
1	28.38%	12.50%	6.67%	95.56%	70.59%
2	28.23%	6.25%	6.67%	95.56%	70.59%
3	28.23%	6.25%	6.67%	95.56%	70.59%
4	28.23%	6.25%	6.45%	95.56%	71.57%
5	28.23%	6.25%	6.45%	95.56%	71.57%
6	28.23%	6.25%	6.67%	95.56%	70.59%
7	28.53%	6.25%	7.10%	95.56%	70.59%
8	27.93%	6.25%	6.24%	95.56%	70.59%
9	28.53%	6.25%	6.88%	95.56%	71.57%
10	28.38%	6.25%	6.67%	95.56%	71.57%
11	28.38%	6.25%	6.67%	95.56%	71.57%
12	28.23%	6.25%	6.67%	95.56%	70.59%

13	28.38%	6.25%	6.88%	95.56%	70.59%
14	28.23%	6.25%	6.45%	95.56%	71.57%
15	28.53%	6.25%	6.88%	95.56%	71.57%
16	28.38%	6.25%	6.67%	95.56%	71.57%
17	28.38%	6.25%	6.88%	95.56%	70.59%
18	28.23%	6.25%	6.67%	95.56%	70.59%
19	28.23%	6.25%	6.45%	95.56%	71.57%
20	28.53%	6.25%	6.88%	95.56%	71.57%

5.4 Optimización de los parámetros y ajuste del RF

Luego de correr la optimización SMBO los parámetros optimizados fueron *mtry* iguales a 17, *min.node.size* igual a 6 y *sample.fraction* igual a 0.43. La exactitud balanceada (BAC) del modelo final fue de 0.57. El error de predicción OOB fue del 26,6%. La convergencia de la BAC se traza en la Figura 2. El número de árboles utilizados (10001) fue más que suficiente para que la BAC se estabilizara. La importancia de la permutación se muestra en la Figura 3. Las dos variables más importantes son VEST y MutPred, predictores que también eran importantes para la puntuación REVEL.

Figura 2 Convergencia de la exactitud balanceada

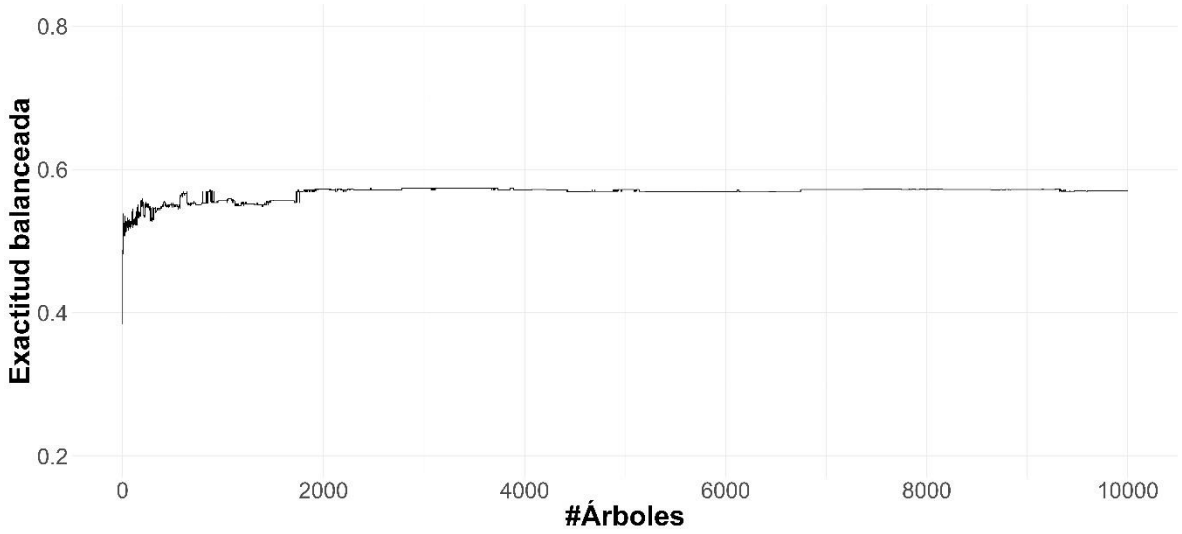
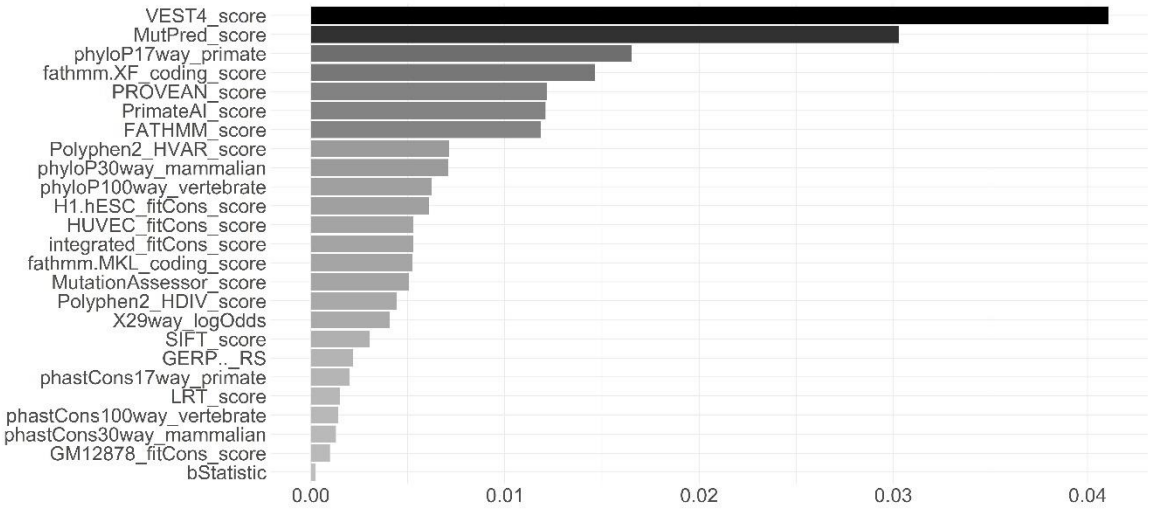


Figura 3 Importancia de las variables por permutaciones



Las probabilidades de pertenencia de grupo en los datos OOB se grafican en la Figura 4. Utilizando estas probabilidades para clasificar variantes, la matriz de confusión OOB se presenta en la Tabla 8 Matriz de confusión OOB del RF. La sensibilidad y la especificidad son altas en la clase benigna. Ninguna de las variantes posiblemente patogénicas se clasifica como benigna. Las otras tres clases son inseparables con estos predictores. La clase posiblemente patogénicas tiene alta sensibilidad y baja especificidad. Las clases patogénicas y patogénicas probables tienen una baja sensibilidad y alta especificidad.

Figura 4 Probabilidades de pertenencia de grupo OOB vs patogenicidad observada

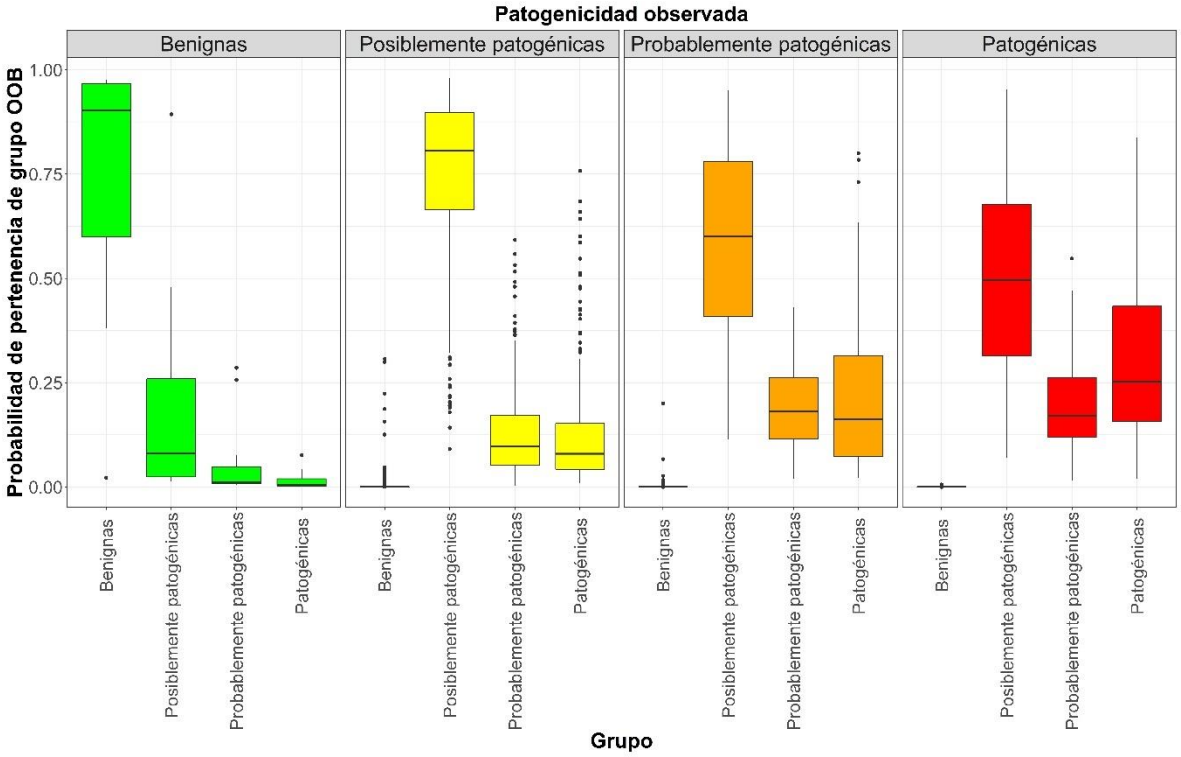


Tabla 8 Matriz de confusión OOB del RF

	Observada				
Predicción	Benignas	Posiblemente patogénicas	Probablemente patogénicas	Patogénicas	**Total**
Benignas	15	0	0	0	15 (2%)
Posiblemente patogénicas	1	439	72	63	575 (87%)
Probablemente patogénicas	0	8	5	4	17 (20%)
Patogénicas	0	18	13	35	66 (10%)
Sensibilidad	0.94	0.94	0.06	0.34	
Especificidad	1	0.35	0.98	0.95	
Valor Pred Pos	1	0.76	0.29	0.53	
Valor Pred Neg	1	0.73	0.87	0.89	

5.5 Refinamientos

Las curvas de características operativas OOB, por clase de patogenicidad se encuentran en la Figura 5. Todas las clases tienen un área bajo la curva por encima de .7, excepto para la clase *Probablemente Patogénicas*, que es la menos distinguible de las tres clases inseparables.

Para TP53MiPaPred_sens el punto de corte para la clase patogénica requerido, para una sensibilidad del 90%, fue 0.07. La Tabla 9 es la matriz de confusión OOB de este modelo. En lugar de solo descartar variantes benignas (2,2% de las variantes), TP53MiPaPred_sens también propone dar prioridad a las variantes predichas como patogénicas (61,5%).

Figura 5 Curvas ROC por clase de patogenicidad

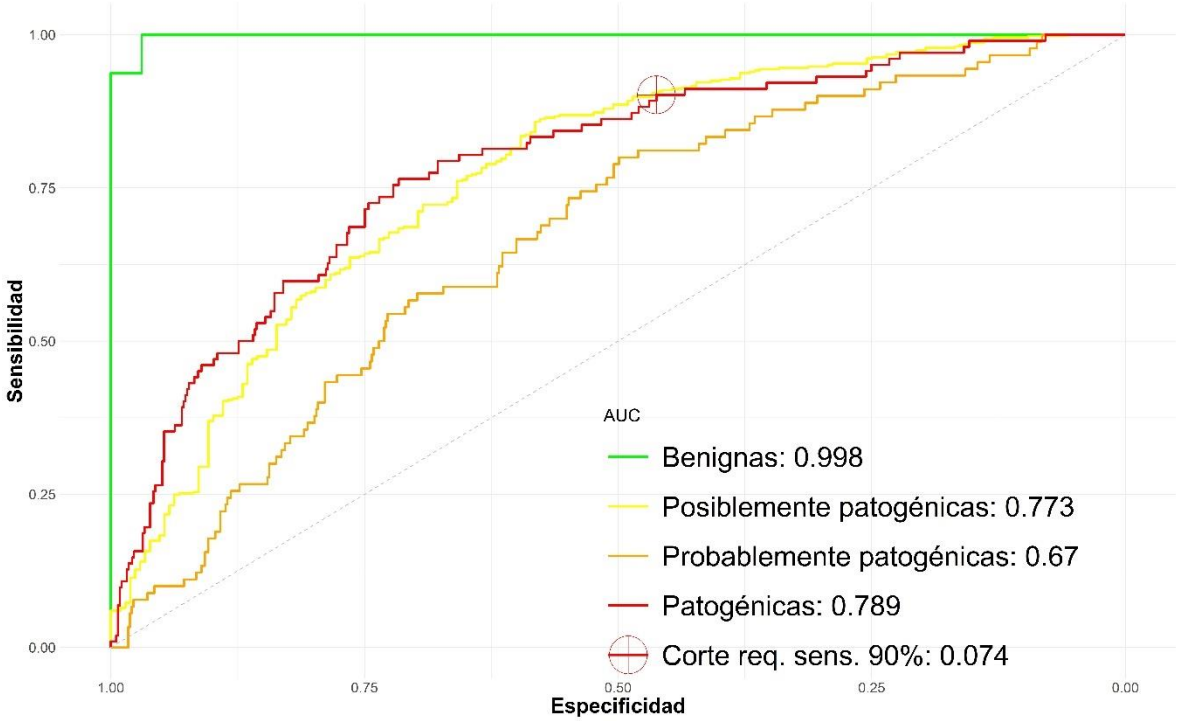


Tabla 9 Matriz de confusión OOB de Tp53MiPaPred_sens

	Observada				
Predicción	Benignas	Posiblemente patogénicas	Probablemente patogénicas	Patogénicas	**Total**
Benignas	15	0	0	0	15 (2.2%)
Posiblemente patogénicas	1	212	21	10	244 (36.3%)
Probablemente patogénicas	0	0	0	0	0 (0.0%)
Patogénicas	0	253	69	92	414 (61.5%)
Sensibilidad	0.94	0.46	0.00	**0.90**	
Especificidad	1.00	0.85	1.00	0.44	
Valor Pred Pos	1.00	0.87	-	0.22	
Valor Pred Neg	1.00	0.41	0.87	0.96	

Figura 6 Curva VPP vs Sensibilidad para la clase Patogénica

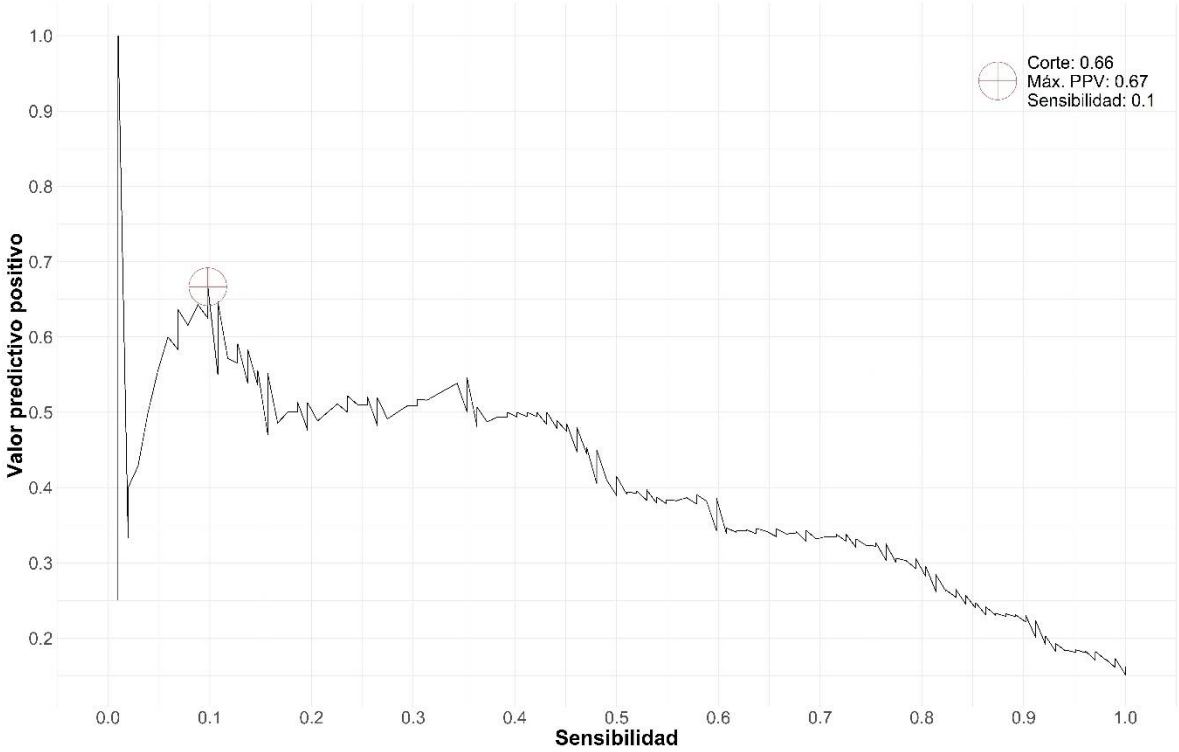


Tabla 10 Matriz de confusión OOB de Tp53MiPaPred_ppv

	Observada				
Predicción	Benignas	Posiblemente patogénicas	Probablemente patogénicas	Patogénicas	**Total**
Benignas	15	0	0	0	15 (2.2%)
Posiblemente patogénicas	1	449	78	77	605 (89.9%)
Probablemente patogénicas	0	14	9	15	38 (5.6%)
Patogénicas	0	2	3	10	15 (2.2%)
Sensibilidad	0.94	0.97	0.10	0.10	
Especificidad	1.00	0.25	0.95	0.99	
Valor Pred Pos	1.00	0.74	0.24	**0.67**	
Valor Pred Neg	1.00	0.76	0.87	0.86	

Después de examinar la curva VPP vs sensibilidad de la clase patogénica (Figura 6), el mejor corte fue de 0,66, correspondiente a un máximo VPP de 0,67. Cada variante con una probabilidad de pertenencia a la clase patogénica por encima de este límite es predicha como patogénica por TP53MiPaPred_ppv. La Tabla 10 es la matriz de confusión OOB de este modelo. Pocas variantes serían predichas como patogénica con este modelo, pero con una confianza de que diez de cada quince serían realmente patogénicas.

5.6 Aplicación de los modelos a las VUS

La Tabla 11 muestra las predicciones TP53MiPaPred sobre las variantes VUS. De 994 variantes VUS, 45(4.5%) se predicen como benignos. Esto es significativamente diferente al 46,7% de variantes predichas como Benignas usando la dicotomización sugerida para REVEL.

Además de descartar estas variantes benignas, TP53MiPaPred_sens priorizar 226(22.7%) como variantes patogénicas. Alrededor del 90% de las verdaderas variantes patogénicas deben estar en este grupo. Sin embargo, cuando se utiliza TP53MiPaPred_ppv ninguna de las VUS se predijo como patogénica. Esto significa que no hubo una sola VUS que pudiera priorizarse como patogénica con alta confianza de serlo.

Tabla 11 Tp53MiPaPred aplicado a variantes de significancia incierta

Predicción	TP53MiPaPred_sens	TP53MiPaPred_ppv
Benignas	45 (4.5%)	45 (4.5%)
Posiblemente patogénicas	723 (72.7%)	942 (94.8%)
Probablemente patogénicas	0 (0.0%)	7 (0.7%)
Patogénicas	226 (22.7%)	0 (0.0%)

5.7 Implementación

Con fines de alcanzar un análisis reproducible, los datos, códigos, el análisis y los resultados han sido cargados en la plataforma Github, bajo la licencia GNU Lesser General Public License v3.0. Por los momentos el repositorio es privado, hasta la aprobación del presente trabajo y se encuentra en el enlace:

https://github.com/peterolejua/Tp53MiPaPred

En este repositorio los usuarios podrán encontrar en la sección de resultados los datos utilizados para el entrenamiento de TP53MiPaPred, junto a las probabilidades in-sample y OOB para cada variante.

En otro conjunto de datos, se encuentran las probabilidades de pertenencia de grupo y la predicción de clase para las variantes de significancia incierta con predictores completos.

Los usuarios interesados en realizar predicciones sobre un nuevo conjunto de variantes deberán contar con los predictores completos o determinar y emplear el método imputación más adecuado para dichas variantes (incluyendo el método de aproximación de RF mediante el RF creado aquí).

Para realizar nuevas predicciones se pone a disposición de los usuarios un archivo "load_to_predict.RData" de R que contiene el RF para predecir las probabilidades de pertenencia de grupo y la función para obtener las clases predichas por TP53MiPaPred_sens y TP53MiPaPred_ppv.

Capítulo 6. Discusión

Entre últimas recomendaciones para la predicción de patogenicidad de las variantes missense TP53 incluyeron el uso del meta-algoritmo REVEL. Usando la base de datos UMD TP53, mostramos que el uso de este algoritmo de una manera dicotomizada sólo era útil para distinguir las variantes benignas en este gen. Además, cuando se aplica a las variantes de significancia incierta, casi la mitad fueron predichas como benignas. Esto es contradictorio con la baja frecuencia de las variantes benignas que se encuentran en este gen. Aunque un rendimiento decente en este gen, REVEL no fue entrenado para ser utilizado específicamente en TP53.

Con REVEL como inspiración, se desarrolló un nuevo meta-algoritmo exclusivamente para las variantes de missense TP53. Para ello se utilizaron las bases de datos UMD TP53 y dbNSFP. En lugar de predecir una patogenicidad asumida, de forma dicotomizada, la nueva herramienta in-silico TP53MiPaPred predice la anotación directa de patogenicidad en cuatro niveles. Se ha entrenado y ajustado utilizando las últimas recomendaciones de la literatura para la optimización y desarrollo de Random Forests. Los datos OOB se utilizaron para la sintonización, para estimar el error fuera de muestra y para un refinamiento post-hoc.

Entre los 25 predictores ensamblados, VEST y MutPred son los más importantes. Sin embargo, con estos predictores, las clases de patogenicidad son inseparables. La clase Benigna es la única distinguible de ellas. Entre las diferentes estrategias para tratar con clases inseparables superpuestas, optamos por la estrategia de reelaborar el problema y realizamos un refinamiento post-hoc basado en curvas ROC y VPP con datos OOB.

Se refinaron dos modelos para predecir mejor la clase patogénica. Ambos son excelentes prediciendo la clase benigna. En términos de priorizar variantes, cualquiera de los dos modelos se puede utilizar para descartar variantes benignas. TP53MiPaPred_sens está destinado a ser utilizado en escenarios donde se necesita una alta sensibilidad, y sólo pocas variantes patogénicas pueden perderse en la priorización. TP53MiPaPred_ppv está destinado a ser utilizado para priorizar pocas variantes con alta probabilidad de ser patógeno. Tp53MiPaPred no se resuelve el problema de cuantificar el impacto funcional que puedan tener las variantes missense de TP53. En cambio, la función del modelo es la de descartar variantes Benignas y priorizar variantes como Patogénicas. Por ejemplo, las variantes de significancia incierta predichas como Patogénicas por TP53MiPaPred pueden ser analizadas en futuros estudios moleculares y poblacionales que confirmen su patogenicidad, impacto clínico (diagnóstico, pronóstico y tratamiento) o su utilidad en investigación de medicamentos o terapias en cáncer.

Desde el punto de vista del análisis empleado, una limitación de este estudio es que, luego de la imputación las estimaciones del error OOB podrían ser pesimistas. Sin embargo, dado la baja cantidad de datos faltantes no es probable que esta sea un sesgo que afecte los resultados del estudio. Por otro lado, no es claro si esto podría mitigarse con el siguiente posible sesgo. Los puntos de corte para TP53MiPaPred_sens y TP53MiPaPred_ppv se eligieron con un refinamiento post-hoc con los datos OOB. Por lo tanto, las matrices de confusión en los datos OOB para estos dos modelos podrían ser ligeramente optimistas. En todo caso, debido a que es necesario realizar futuros estudios de validación externa con conjuntos independientes de variantes, estos posibles sesgos podrían analizarse más adelante.

La presente propuesta se limitó a la exploración de la técnica de Random Forest. No se abarcaron técnicas de aprendizaje profundo o boosting. El supuesto es que, como se menciona arriba, hasta ahora las técnicas de Random Forest han tenido un poder predictivo equiparable. Queda entonces como recomendación para futuros estudios la exploración de estas técnicas en el problema específico tomando el desenlace como ordinal. Otras técnicas clásicas como regresión u otras técnicas de machine learning han demostrado tener un menor poder predictivo que Random Forest (13,48,102,126).

Posibles sesgos podrían surgir de una mala clasificación del desenlace de patogenicidad. Sin embargo, no existe un consenso único establecido para la clasificación de patogenicidad. En este trabajo para el caso particular de las variantes missense de TP53, hemos tomado la patogenicidad ya curada en la base de datos de UMD TP53 (6). Esta patogenicidad se ha establecido siguiendo adaptaciones que se han hecho de las guías de la ACMG-AMP. Una limitante de este estudio es que dichas adaptaciones aún se encuentran en elaboración para su publicación (18). En todo caso con el meta-predictor propuesto se ha evidenciado una manera de priorizar variantes. El esquema de trabajo es reproducible al desarrollo de futuras herramientas in-silico para este y otros genes.

Es necesario desarrollar más herramientas in-silico específicas para TP53. El reajuste de algoritmos antiguos para estas variantes (no meta-algoritmos) es una posibilidad, especialmente aquellos que se fueron más importantes para TP53MiPaPred. Se podría utilizar una mezcla de predictores que se encuentran en diferentes herramientas individuales. Hasta entonces, el algoritmo aquí propuesto podría integrarse en bases de datos como el UMD TP53 o dbNSFP y podría mostrarse en plataformas como cBioPortal al analizar variantes de missense TP53.

Hasta el momento se recomienda el uso en combinación de los algoritmos Align-GVGD y BayesDel por ser un poco mejor que la combinación Align-GVGD y REVEL. En próximos estudios se podría evaluar la combinación Align-GVGD y TP53MiPaPred.

Capítulo 7. Conclusiones

En la búsqueda de mejores modelos predictivos para la patogenicidad de variantes genéticas, aquí se propuso un meta-predictor para ser utilizado específicamente con las variantes missense del gen TP53. En problemas de priorización, la nueva herramienta in-silico puede usarse para descartar variantes benignas y, con dos enfoques diferentes, para priorizar variantes que podrían ser patogénicas. Futuros estudios deberán evaluar la posible implementación del modelo en combinación con otras herramientas in-silico utilizadas en la literatura. Potencialmente, TP53MiPaPred podría formar parte de adaptaciones hechas a guías para la interpretación de variantes de TP53. Para la implementación del modelo se creó un repositorio en Github con fines de investigación reproducible y que pone a disposición de los usuarios la estructura de trabajo, así como las predicciones del modelo. Este estará públicamente disponible en el enlace https://github.com/peterolejua/Tp53MiPaPred, previa aprobación del presente trabajo.

Referencias

1. Dolgin E. The most popular genes in the human genome. Nature. 2017 23;551(7681):427–31.

2. Hainaut P, Pfeifer GP. Somatic TP53 Mutations in the Era of Genome Sequencing. Cold Spring Harb Perspect Med [Internet]. 2016 Nov [cited 2019 Jul 31];6(11). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5088513/

3. Leroy B, Ballinger ML, Baran-Marszak F, Bond GL, Braithwaite A, Concin N, et al. Recommended Guidelines for Validation, Quality Control, and Reporting of *TP53* Variants in Clinical Practice. Cancer Res [Internet]. 2017 Mar 2 [cited 2018 May 14]; Available from: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-16-2179

4. Boettcher S, Miller PG, Sharma R, McConkey M, Leventhal M, Krivtsov AV, et al. A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. Science. 2019 Aug 9;365(6453):599–604.

5. Hattori H. [Li-Fraumeni Syndrome-Current Status and Prospects in Clinical Practice]. Gan To Kagaku Ryoho. 2019 Jul;46(7):1103–8.

6. Malcikova J, Tausch E, Rossi D, Sutton LA, Soussi T, Zenz T, et al. ERIC recommendations for TP53 mutation analysis in chronic lymphocytic leukemia-update on methodological approaches and results interpretation. Leukemia. 2018 May;32(5):1070–80.

7. Li VD, Li KH, Li JT. TP53 mutations as potential prognostic markers for specific cancers: analysis of data from The Cancer Genome Atlas and the International Agency for Research on Cancer TP53 Database. J Cancer Res Clin Oncol. 2019 Mar;145(3):625–36.

8. Wu C-H, Hwang M-J. Risk stratification for lung adenocarcinoma on EGFR and TP53 mutation status, chemotherapy, and PD-L1 immunotherapy. Cancer Med. 2019 Aug 13;

9. Williams DS, Mouradov D, Browne C, Palmieri M, Elliott MJ, Nightingale R, et al. Overexpression of TP53 protein is associated with the lack of adjuvant chemotherapy benefit in patients with stage III colorectal cancer. Mod Pathol Off J U S Can Acad Pathol Inc. 2019 Aug 30;

10. Bittar CM, Vieira IA, Sabato CS, Andreis TF, Alemar B, Artigalás O, et al. TP53 variants of uncertain significance: increasing challenges in variant interpretation and genetic counseling. Fam Cancer. 2019 Jul 18;

11. Jimenez Rodriguez B, Diaz Córdoba G, Garrido Aranda A, Álvarez M, Vicioso L, Llácer Pérez C, et al. Detection of TP53 and PIK3CA Mutations in Circulating Tumor DNA Using Next-Generation Sequencing in the Screening Process for Early Breast Cancer Diagnosis. J Clin Med. 2019 Aug;8(8):1183.

12. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014 Apr;508(7497):469–76.

13. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol. 2017 Nov 28;18(1):225.

14. Rivera-Muñoz EA, Milko LV, Harrison SM, Azzariti DR, Kurtz CL, Lee K, et al. ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. Hum Mutat. 2018;39(11):1614–22.

15. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho Y-Y, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet Med Off J Am Coll Med Genet. 2017 Oct;19(10):1105–17.

16. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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 Off J Am Coll Med Genet. 2015 May;17(5):405–24.

17. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. J Mol Diagn. 2017 Jan;19(1):4–23.

18. Tikkanen T, Leroy B, Fournier JL, Risques RA, Malcikova J, Soussi T. Seshat: A Web service for accurate annotation, validation, and analysis of *TP53* variants

generated by conventional and next-generation sequencing. Hum Mutat. 2018 Jul;39(7):925–33.

19. Fortuno C, Cipponi A, Ballinger ML, Tavtigian SV, Olivier M, Ruparel V, et al. A quantitative model to predict pathogenicity of missense variants in the TP53 gene. Hum Mutat. 2019;40(6):788–800.

20. Bean LJH, Hegde MR. Clinical implications and considerations for evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Med [Internet]. 2017 Dec 18 [cited 2018 May 4];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5733812/

21. Fortuno C, James PA, Young EL, Feng B, Olivier M, Pesaran T, et al. Improved, ACMG-compliant, in silico prediction of pathogenicity for missense substitutions encoded by TP53 variants. Hum Mutat. 2018 Aug;39(8):1061–9.

22. Collins H, Calvo S, Greenberg K, Forman Neall L, Morrison S. Information Needs in the Precision Medicine Era: How Genetics Home Reference Can Help. Interact J Med Res. 2016 Apr 27;5(2):e13.

23. Reference GH. What is precision medicine? [Internet]. Genetics Home Reference. [cited 2019 Feb 7]. Available from: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

24. What is Precision Medicine? [Internet]. [cited 2019 Feb 7]. Available from: https://learn.genetics.utah.edu/content/precision/intro/

25. Genomics and Future of Medicine- Business News [Internet]. [cited 2019 Feb 7]. Available from: https://www.businesstoday.in/magazine/features/genomics-and-future-of-medicine/story/243003.html

26. Huntington W. The Human Genome: A Window on Human Genetics, Biology, and Medicine. In: Genomic and Personalized Medicine. 2014. p. 5.

27. The Cost of Sequencing a Human Genome [Internet]. National Human Genome Research Institute (NHGRI). [cited 2019 Feb 7]. Available from: https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

28. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med Off J Am Coll Med Genet. 2018;20(10):1122–30.

29. Panda B, Krishman N. BIOINFORMATICS, SYSTEMS BIOLOGY, AND SYSTEMS MEDICINE. In: Kumar D, Eng C, editors. Genomic medicine: principles and practice. Second edition. Oxford: Oxford University Press; 2015.

30. Richards JE, Hawley RS, Hawley RS. The human genome: a user's guide. 3rd ed. Amsterdam Boston: Elsevier Academic Press; 2011. 585 p.

31. Kato S, Han S-Y, Liu W, Otsuka K, Shibata H, Kanamaru R, et al. Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc Natl Acad Sci U S A. 2003 Jul 8;100(14):8424–9.

32. Id Said B, Kim H, Tran J, Novokmet A, Malkin D. Super-Transactivation TP53 Variant in the Germline of a Family with Li-Fraumeni Syndrome. Hum Mutat. 2016;37(9):889–92.

33. Pejaver V, Mooney SD, Radivojac P. Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. Hum Mutat. 2017 Sep;38(9):1092–1108.

34. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, et al. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. Mol Cell. 2018 05;71(1):178-190.e8.

35. ; on behalf of the ACMG Laboratory Quality Assurance Committee, Richards S, Aziz N, Bale S, Bick D, Das S, 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 May;17(5):405–23.

36. Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, et al. Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016 Jun 2;98(6):1067–76.

37. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian

classification framework. Genet Med [Internet]. 2018 Jan 4 [cited 2018 May 4]; Available from: https://www.nature.com/articles/gim2017210

38. A centuries-old math equation used to solve a modern-day genetics challenge [Internet]. ScienceDaily. [cited 2018 May 17]. Available from: https://www.sciencedaily.com/releases/2018/01/180118100802.htm

39. TP53 Variant Curation Expert Panel. TP53 Rule Specifications for the ACMG/AMP Variant Curation Guidelines [Internet]. 2019 [cited 2019 Oct 16]. Available from: https://clinicalgenome.org/site/assets/files/3876/clingen_tp53_acmg_specifications_v1.pdf

40. Mester JL, Ghosh R, Pesaran T, Huether R, Karam R, Hruska KS, et al. Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel. Hum Mutat. 2018 Nov;39(11):1581–92.

41. Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Hum Mutat. 2015 May;36(5):513–523.

42. Chen Q, Dai C, Zhang Q, Du J, Li W. Evaluation of performance of five bioinformatics software for the prediction of missense mutations. Zhonghua Yi Xue Yi Chuan Xue Za Zhi Zhonghua Yixue Yichuanxue Zazhi Chin J Med Genet. 2016 Oct;33(5):625–8.

43. Ernst C, Hahnen E, Engel C, Nothnagel M, Weber J, Schmutzler RK, et al. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. BMC Med Genomics. 2018 Mar;11(1):35.

44. Crockett DK, Lyon E, Williams MS, Narus SP, Facelli JC, Mitchell JA. Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants. J Am Med Inform Assoc JAMIA. 2012;19(2):207–211.

45. Li Q, Liu X, Gibbs RA, Boerwinkle E, Polychronakos C, Qu H-Q. Gene-specific function prediction for non-synonymous mutations in monogenic diabetes genes. PloS One. 2014;9(8):e104452.

46. Wang M, Wei L. iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. Sci Rep. 2016 Aug;6:31321.

47. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct;99(4):877–85.

48. Qi H, Chen C, Zhang H, Long JJ, Chung WK, Guan Y, et al. MVP: predicting pathogenicity of missense variants by deep learning. 2018 Apr 2 [cited 2018 May 17]; Available from: http://biorxiv.org/lookup/doi/10.1101/259390

49. Chen L, Qin ZS. Using DIVAN to assess disease/trait-associated single nucleotide variants in genome-wide scale. BMC Res Notes [Internet]. 2017 Dec [cited 2018

May 17];10(1). Available from: http://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-017-2851-y

50. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–1081.

51. Flanagan SE, Patch A-M, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomark. 2010 Aug;14(4):533–537.

52. Capriotti E, Martelli PL, Fariselli P, Casadio R. Blind prediction of deleterious amino acid variations with SNPs&GO: CAPRIOTTI et al. Hum Mutat. 2017 Sep;38(9):1064–71.

53. Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. Genetics. 2016 Jun;203(2):635–47.

54. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. 2017 Feb;100(2):267–80.

55. Ravichandran V, Shameer Z, Kemel Y, Walsh M, Cadoo K, Lipkin S, et al. Towards automation of germline variant curation in clinical cancer genetics. bioRxiv. 2018 May 1;295865.

56. The TP53 Website - Seshat [Internet]. [cited 2018 May 17]. Available from: https://p53.fr/tp53-database/seshat

57. Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.

58. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018 Oct 4;103(4):474–83.

59. Iddamalgoda L, Das PS, Aponso A, Sundararajan VS, Suravajhala P, Valadi JK. Data Mining and Pattern Recognition Models for Identifying Inherited Diseases: Challenges and Implications. Front Genet [Internet]. 2016 Aug 10 [cited 2018 May 18];7. Available from: http://journal.frontiersin.org/Article/10.3389/fgene.2016.00136/abstract

60. Probst P, Boulesteix A-L. To Tune or Not to Tune the Number of Trees in Random Forest. J Mach Learn Res. 2018;18(181):1–18.

61. Hornung R. Ordinal Forests. J Classif [Internet]. 2019 Jan 22 [cited 2019 Oct 17]; Available from: https://doi.org/10.1007/s00357-018-9302-x

62. Janitza S, Tutz G, Boulesteix A-L. Random forest for ordinal responses: Prediction and variable selection. Comput Stat Data Anal. 2016 Apr;96:57–73.

63. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017 Mar 31;77(1):1–17.

64. Kruppa J, Liu Y, Biau G, Kohler M, König IR, Malley JD, et al. Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. Biom J. 2014;56(4):534–63.

65. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability Machines: Consistent Probability Estimation Using Nonparametric Learning Machines. Methods Inf Med. 2012;51(01):74–81.

66. Janitza S, Hornung R. On the overestimation of random forest's out-of-bag error. PloS One. 2018;13(8):e0201904.

67. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. In: 2010 20th International Conference on Pattern Recognition [Internet]. Istanbul, Turkey: IEEE; 2010 [cited 2019 Oct 18]. p. 3121–4. Available from: http://ieeexplore.ieee.org/document/5597285/

68. Brzezinski D, Stefanowski J, Susmaga R, Szczech I. On the Dynamics of Classification Measures for Imbalanced and Streaming Data. IEEE Trans Neural Netw Learn Syst. 2019;1–11.

69. García V, Mollineda RA, Sánchez JS. Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions. In: Araujo H, Mendonça AM, Pinho AJ, Torres MI, editors. Pattern Recognition and Image Analysis. Springer Berlin Heidelberg; 2009. p. 441–8. (Lecture Notes in Computer Science).

70. Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip Rev Data Min Knowl Discov. 2019;9(3):e1301.

71. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min. 2017 Dec;10(6):363–77.

72. Sage A. Random forest robustness, variable importance, and tree aggregation [Internet] [Doctor of Philosophy]. [Ames]: Iowa State University, Digital Repository; 2018 [cited 2019 Oct 20]. Available from: https://lib.dr.iastate.edu/etd/16453/

73. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006 Jun 1;27(8):861–74.

74. Ozenne B, Subtil F, Maucort-Boulch D. The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol. 2015 Aug 1;68(8):855–9.

75. Xiong H, Wu J, Liu L. Classification with ClassOverlapping: A Systematic Study. In: Proceedings of the 1st International Conference on E-Business Intelligence (ICEBI2010) [Internet]. Atlantis Press; 2010. Available from: https://doi.org/10.2991/icebi.2010.43

76. Lin W-J, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform. 2013 Jan 1;14(1):13–26.

77. Prati RC, Batista GEAPA, Monard MC. Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior. In: Monroy R, Arroyo-Figueroa G, Sucar LE, Sossa H, editors. MICAI 2004: Advances in Artificial Intelligence. Springer Berlin Heidelberg; 2004. p. 312–21. (Lecture Notes in Computer Science).

78. Santiso S, Casillas A, Pérez A. The class imbalance problem detecting adverse drug reactions in electronic health records. Health Informatics J. 2018 Sep 19;1460458218799470.

79. Sáez JA, Galar M, Krawczyk B. Addressing the Overlapping Data Problem in Classification Using the One-vs-One Decomposition Strategy. IEEE Access. 2019;7:83396–411.

80. Rocca B. Handling imbalanced datasets in machine learning [Internet]. Medium. 2019 [cited 2019 Oct 9]. Available from: https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28

81. Mahmood K, Jung C-H, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics. 2017 May;11(1):10.

82. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018 Aug;50(8):1161.

83. Mirón Canelo JA, Alonso Sardón M, Iglesias de Sena H. Metodología de investigación en Salud Laboral. Med Segur Trab. 2010 Dec;56(221):347–65.

84. Ruiz Morales Á de J, Camacho J, Delgado Ramírez MB, Dennis Verano RJ, Duarte Osorio A, Gómez-Restrepo C, et al. Epidemiología clínica investigación clínica aplicada. ///.

85. Types of genetic variation studies [Internet]. EMBL-EBI Train online. 2017 [cited 2019 Feb 28]. Available from: https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction/types-genetic-variation-studies

86. Studies on the functional consequences of variants [Internet]. EMBL-EBI Train online. 2017 [cited 2019 Feb 28]. Available from: https://www.ebi.ac.uk/training/online/course/human-genetic-variation-introduction/what-genetic-variation/phenotypes-mendelian-and-0

87. Soussi T, Leroy B, Taschner PEM. Recommendations for analyzing and reporting TP53 gene variants in the high-throughput sequencing era. Hum Mutat. 2014 Jun;35(6):766–78.

88. Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, et al. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. Hum Mutat. 2016;37(9):865–76.

89. Leroy B, Anderson M, Soussi T. TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum Mutat. 2014 Jun;35(6):672–88.

90. Soussi T, Asselain B, Hamroun D, Kato S, Ishioka C, Claustres M, et al. Meta-analysis of the p53 mutation database for mutant p53 biological activity reveals a methodologic bias in mutation detection. Clin Cancer Res Off J Am Assoc Cancer Res. 2006 Jan 1;12(1):62–9.

91. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat. 2016 Mar;37(3):235–41.

92. dbNSFP - Jpopgen [Internet]. [cited 2019 Feb 25]. Available from: https://sites.google.com/site/jpopgen/dbNSFP

93. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat. 2008 Jan;29(1):6–13.

94. dbNSFP4.0b1a.readme.txt [Internet]. Google Docs. [cited 2019 Feb 25]. Available from: https://drive.google.com/file/d/1neHesnuX1RbSOxfXvbpJGwS3xSEdYwUG/view?usp=sharing&usp=embed_facebook

95. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Vol. 11. 2014. 361–362 p.

96. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015 Apr;24(8):2125–2137.

97. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–6.

98. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. 2017 [cited 2019 Feb 26]; Available from: http://europepmc.org/abstract/ppr/ppr28042

99. Raimondi D, Tanyalcin I, Ferte J, Gazzo A, Orlando G, Lenaerts T, et al. DEOGEN2: prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017 Jul;45(W1):W201–W206.

100. Balasubramanian S, Fu Y, Pawashe M, McGillivray P, Jin M, Liu J, et al. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. Nat Commun. 2017 Aug 29;8(1):382.

101. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 8;47(D1):D886–94.

102. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinforma Oxf Engl. 2015 Mar;31(5):761–763.

103. IONITA-LAZA I, MCCALLUM K, XU B, BUXBAUM J. A SPECTRAL APPROACH INTEGRATING FUNCTIONAL GENOMIC ANNOTATIONS FOR CODING AND NONCODING VARIANTS. Nat Genet. 2016 Feb;48(2):214–20.

104. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated

Analysis of Annotation Data. Sci Rep [Internet]. 2015 Jun 30 [cited 2019 Feb 27];5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4444969/

105. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017 Apr;49(4):618–624.

106. Mutalyzer 2.0.29 — Batch Job Interface [Internet]. [cited 2019 Mar 4]. Available from: https://mutalyzer.nl/batch-jobs?job_type=position-converter

107. Gene: TP53 (ENSG00000141510) - Summary - Homo sapiens - Ensembl genome browser 95 [Internet]. [cited 2019 Mar 4]. Available from: https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000141510;r=17:7668402-7687538;t=ENST00000269305

108. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

109. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

110. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009 Jan 9;19(9):1553–61.

111. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011 Sep;39(17):e118.

112. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Hum Mutat. 2013 Jan;34(1):57–65.

113. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinforma Oxf Engl. 2015 May 15;31(10):1536–43.

114. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinforma Oxf Engl. 2018 Feb;34(3):511–513.

115. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE [Internet]. 2012 Oct 8 [cited 2019 Feb 26];7(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3466303/

116. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3:S3.

117. Douville C, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, et al. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). Hum Mutat. 2016 Jan;37(1):28–35.

118. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam H, et al. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. 2017 [cited 2019 Feb 26]; Available from: http://europepmc.org/abstract/ppr/ppr28539

119. Gulko B, Hubisz MJ, Gronau I, Siepel A. Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. Nat Genet. 2015 Mar;47(3):276–83.

120. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol [Internet]. 2010 Dec 2 [cited 2019 Feb 27];6(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996323/

121. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan;20(1):110–21.

122. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Jan 8;15(8):1034–50.

123. McVicker G, Gordon D, Davis C, Green P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. PLOS Genet. 2009 May 8;5(5):e1000471.

124. Hastie T, Tibshirani R, Friedman J. Model Assessment and Selection. In: The Elements of Statistical Learning [Internet]. 2nd ed. New York, NY: Springer New York; 2009 [cited 2019 Oct 9]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-84858-7

125. R: The R Project for Statistical Computing [Internet]. [cited 2019 Mar 5]. Available from: https://www.r-project.org/

126. Pan Y, Liu D, Deng L. Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. PloS One. 2017;12(6):e0179314.