UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

“CARACTERIZACIÓN DE LA DIVERSIDAD EN GENES IMPORTANTES DE FARMACOGENÓMICA MEDIANTE SECUENCIACIÓN DE NUEVA GENERACIÓN”

TESIS

QUE PARA OPTAR POR EL GRADO DE:

DOCTOR EN CIENCIAS

PRESENTA:
OMAR FERNANDO CRUZ CORREA

TUTOR PRINCIPAL:
Dr. Francisco Xavier Soberón Mainero — Instituto de Biotecnología, UNAM

MIEMBROS DEL COMITÉ TUTOR:
Dr. Ruy López Ridaura — Instituto Nacional de Salud Pública
Dr. José Pedraza Chaverri — Facultad de Química, UNAM

CIUDAD DE MÉXICO, ENERO, 2019

CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Significado de abreviaturas

AMI	Residentes de comunidades indígenas mexicanas, genotipificados en el INMEGEN
AUC	Área bajo la curva de concentración plasmática de atorvastatina desde el tiempo cero hasta la última medición (48 horas)
ATV	Atorvastatina
CEU	Residentes de Utah, Estados Unidos, con ascendencia europea, genotipificados por el proyecto de los 1000 genomas
FATHMM-MKL	Functional Analysis through Hidden Markov Models – Multiple Kernel Learning
HMG-CoA	3-hidroxi-3-metilglutaril coenzima A
IBS	Residentes de España, con ascendencia ibérica, genotipificados por el proyecto de los 1000 genomas
Kaviar	~Known VARiants
LWK	Residentes de Webuye, Kenya con ascendencia luhya, genotipificados por el proyecto de los 1000 genomas
MXL	Residentes de Los Angeles, Estados Unidos con ascendencia mexicana, genotipificados por el proyecto de los 1000 genomas
PharmGKB	Pharmacogenomics Knowledgebase
PROVEAN	Protein Variation Effect Analyzer
RI	Rango Intercuartílico
SIGMA	Iniciativa Slim en Medicina Genómica para las Américas (por sus siglas en inglés)
SNVs	Variantes de un solo nucleótido
MNVs	Variantes de múltiples nucleótidos
TSI	Residentes de Italia con ascendencia toscana, genotipificados por el proyecto de los 1000 genomas
UDP	Uridina difosfato
VEP	Variant Effect Predictor
YRI	Residentes de Ibadan, Nigeria con ascendencia yoruba, genotipificados por el proyecto de los 1000 genomas

AGRADECIMIENTOS

Quiero agradecer al CONACYT por el apoyo económico que me brindó personalmente a través de la beca 366725/245614 para estudios de posgrado y a través de su Convocatoria de Investigación Científica Básica apoyo #252952 “Diversidad Fármacogenética en Mexicanos, colección e interpretación” los cuales hicieron posible la realización de esta investigación.

De la misma manera agradezco al Programa de Apoyo a los Estudios de Posgrado (PAEP) por el apoyo que me permitió presentar los resultados de este trabajo en el congreso internacional “14th International Symposium on Variants in the Genome: detection, sequencing & interpretation” realizado en la ciudad de Santiago de Compostela, España.

Quiero agradecer a mi tutor de maestría y doctorado el doctor Xavier Soberón por toda la confianza que ha depositado en mí y en mi trabajo a lo largo de estos años. Siempre tendré en cuenta todo lo que he aprendido de él.

También a todos mis amigos y compañeros en el Instituto Nacional de Medicina Genómica (INMEGEN), por compartir su tiempo conmigo y ofrecerme su amistad, en especial a los miembros de la Unidad de Secuenciación con quienes trabajé también para realizar la secuenciación de este proyecto: Alfredo Mendoza, Julio Canseco, Salvador Hernández y Haydee Miranda.

A mis padres y a mi hermano les agradezco ser siempre una inspiración que me impulsa a buscar mi propia superación y el cumplimiento de mis metas. También agradezco a Lourdes Ramírez Hobak quien me ha acompañado en la travesía por el doctorado y todos los demás lugares que hemos tenido la suerte de conocer juntos. Espero que me acompañes siempre a pesar de todas las dificultades.

A todos aquellos amigos, compañeros y profesores que de una manera u otra me ofrecieron su consejo, apoyo y amistad e hicieron posible la realización de este trabajo:
Gracias.

1. RESUMEN

Objetivos: Explorar la utilidad de la colección de la diversidad genética en genes involucrados en el metabolismo y respuesta de diversos fármacos en muestras de población mexicana, mediante el modelado de la farmacocinética de atorvastatina en 60 individuos. Asimismo, iniciar la recopilación de un catálogo de las variantes genéticas presentes en genes relacionados a medicamentos para informar subsecuentes estudios de farmacogenómica en mexicanos.

Sujetos y Métodos: Secuenciamos un grupo de 430 genes en 60 voluntarios sanos del noreste de México caracterizados fenotípicamente en cuanto a la farmacocinética de atorvastatina. Construimos un modelo de regresión con vectores de soporte utilizando las variantes presentes en 20 genes que intervienen en el metabolismo y respuesta de atorvastatina y posteriormente lo refinamos para explicar la mayor parte de la variabilidad en AUC de atorvastatina en estos individuos. Las variantes genéticas encontradas en los 60 individuos antes mencionados constituyen un primer catálogo de la diversidad genética en los genes que intervienen en el metabolismo y respuesta de diversos medicamentos. Para complementar este catálogo se incluyó además la información de las variantes encontradas en las mismas regiones genómicas en un grupo de 94 genomas completos de individuos amerindios y 968 exomas completos de individuos mexicanos provenientes de otros proyectos de secuenciación.

Resultados: Los modelos de regresión construidos a partir de las variantes genéticas con efectos funcionales predichos o reportados que se encuentran presentes en genes relacionados con atorvastatina explican una mayor proporción de la varianza que modelos construidos a partir de otros grupos de variantes. El modelo de regresión final refinado utiliza 60 variantes genéticas para explicar el 93.53% de la varianza en AUC de atorvastatina en los 60 individuos. Además se recopiló un catálogo de más de 47 mil variantes genéticas en genes relacionados a diversos medicamentos, de las cuales el 25.49% tienen efectos funcionales predichos o reportados y podrían afectar el metabolismo y respuesta a fármacos en la población mexicana.

Conclusiones: Los resultados soportan el uso de variantes presentes en diversos genes que intervienen en diferentes etapas del metabolismo de un mismo fármaco para predecir un fenotipo de interés, en este caso la farmacocinética de atorvastatina. Además, como consecuencia de este estudio se ha obtenido una lista de variantes con posibles efectos funcionales que pueden afectar el metabolismo y respuesta a fármacos, la cual puede utilizarse para el diseño preliminar de plataformas de genotipificación masiva para su uso en estudios de farmacogenómica en México.

2. ABSTRACT

Aim: To explore the usefulness of genetic diversity collection in genes involved in drug metabolism and response in Mexican populations through the modeling of atorvastatin pharmacokinetics in 60 individuals. As well as to start the compilation of a catalog of genetic variants present in drug related genes to inform subsequent pharmacogenomic studies in Mexicans.

Subjects and methods: We sequenced a group of 430 genes in 60 healthy volunteers from northeastern Mexico phenotypically characterized as to atorvastatin pharmacokinetics. We constructed a support vector regression model using variants present in 20 genes involved in atorvastatin metabolism and response and subsequently refined it to explain most of the AUC variability in these individuals. The genetic variants found in the 60 individuals mentioned above constitute a first catalog of genetic diversity in genes involved in the metabolism and response of various drugs. To complement this catalog we also included information on variants found in the same genomic regions in a group of 94 complete genomes of Amerindian individuals and 968 complete exomes of Mexican individuals from other sequencing projects.

Results: Regression models constructed from genetic variants with predicted or reported functional effects found in atorvastatin-related genes account for a greater proportion of variance than models constructed from other groups of variants. The refined regression model uses 60 genetic variants to account for 93.53% of the variance in atorvastatin AUC in the 60 individuals. Additionally, a catalog of more than 47,000 genetic variants in genes related to various drugs was compiled, 25.49% of these variants have predicted or reported functional effects and could affect drug metabolism and response in Mexican populations.

Conclusions: Our results support the use of variants present in several genes involved in different stages of metabolism of the same drug to predict a specific phenotype, in this case atorvastatin pharmacokinetics. In addition, as a result of this study, we obtained a list of variants with possible functional effects that may affect drug metabolism and response, which can be used for the preliminary design of genotyping platforms useful for pharmacogenomic studies in Mexico.

3. INTRODUCCIÓN

En la actualidad existen guías farmacogenéticas para más de 40 medicamentos que recomiendan la manera de ajustar el tratamiento farmacológico según el genotipo de un individuo. Sin embargo, hasta ahora no ha sido posible desarrollar modelos para algunos fármacos importantes en México. En este trabajo proponemos la utilización de variantes presentes en diversos genes que intervienen en diferentes etapas del metabolismo y respuesta de atorvastatina para modelar su perfil farmacocinético. Asimismo afirmamos la importancia de recopilar un catálogo de las variantes genéticas localizadas en genes relacionados a diversos medicamentos en individuos pertenecientes a la población mexicana para informar estudios subsecuentes de farmacogenómica.

El desarrollo de la farmacogenómica permitiría alcanzar la máxima eficacia terapéutica y evitar la aparición de efectos adversos, beneficiando de forma importante los sistemas de salud en México. En el capítulo de justificación se muestra la existencia de variantes genéticas propias de cada población, las cuales deben ser tomadas en cuenta para explicar o predecir las diferencias interpersonales en cuanto a metabolismo y respuesta a medicamentos.

Posteriormente dentro del marco teórico se señalan las características de la atorvastatina y la importancia de anticipar el perfil farmacocinético de este fármaco en base al genotipo para evitar la aparición de miopatía asociada a su uso. También se mencionan los genes que intervienen en su transporte, metabolismo y respuesta, así como las variantes genéticas que pueden modificar sus efectos en el organismo. Esta información es indispensable ya que el propósito de este trabajo es relacionar el perfil farmacocinético de este fármaco con las variantes genéticas presentes en los genes relacionados con atorvastatina.

En la sección de antecedentes señalamos que en México, hasta el momento no se ha estudiado de forma integral y simultánea todos los genes que se han reportado como asociados a fármacos y que esto puede lograrse gracias a las nuevas tecnologías de secuenciación masiva y los enfoques de enriquecimiento y captura de regiones específicas del genoma que permiten de forma simultánea la evaluación de las variantes genéticas conocidas y la identificación de nuevas variantes potencialmente importantes. Estos enfoques metodológicos permitirán durante el desarrollo de la primera parte experimental de este trabajo catalogar la diversidad genética en un grupo de 430 genes relacionados con el metabolismo y respuesta a fármacos.

En la siguiente sección del trabajo se detallan las características de la metodología seguida: los sujetos analizados, los métodos utilizados para la obtención de los datos de secuenciación y su tratamiento, la predicción del efecto funcional de las variantes genéticas encontradas y la posterior generación de un modelo de regresión para la farmacocinética de atorvastatina utilizando las variantes con mejor predicción funcional que se encuentran en los genes relacionados con el metabolismo y respuesta de este fármaco.

Posteriormente se señalan los resultados más representativos de este trabajo en cuanto a la generación de un modelo de regresión para la farmacocinética de atorvastatina en 60 individuos, así como también en cuanto a la generación de un catálogo de las variantes genéticas presentes en 430 genes relacionados con el metabolismo y respuesta a diversos fármacos.

Después se realiza una discusión de los resultados obtenidos y se resalta como conclusiones que estos resultados soportan el uso de una combinación de herramientas computacionales para predecir los efectos funcionales de las variantes no caracterizadas experimentalmente. Este proceso constituye una primera evaluación de la importancia de las variantes novedosas encontradas mediante esquemas de secuenciación de nueva generación. Además se resalta que el análisis simultáneo e integral de los genes relacionados a diferentes etapas del metabolismo de fármacos puede ayudar a construir modelos que resulten más útiles que los existentes actualmente para explicar la variabilidad interpersonal en la farmacocinética de un fármaco específico.

Por último se comenta sobre las implicaciones potenciales de este trabajo para posteriores estudios de farmacogenómica en la población mexicana y la viabilidad de utilizar herramientas de genotipificación masiva para evaluar la respuesta a fármacos en un ambiente clínico actual y futuro.

4. Planteamiento del problema y justificación

Muchos de los medicamentos utilizados en el tratamiento de diversos padecimientos presentan una tasa de respuesta con valores entre el 50% y 75% (1). Esto implica un hecho fundamental que debe ser reconocido: Debido a la variabilidad interpersonal un porcentaje significativo de los individuos a los que se administre un fármaco no responderán al tratamiento, e incluso, algunos de ellos presentarán reacciones adversas severas que pueden llegar a ser fatales. Por ejemplo, se ha reportado una incidencia de 9.53% de reacciones adversas en pacientes hospitalizados, de las cuales el 12.3% pusieron en peligro la vida de los pacientes (2).

La diferencia observable en el metabolismo y eficacia de los fármacos dependen de una gran variedad de factores, tanto individuales como ambientales. De forma general se estima que los factores genéticos son responsables de aproximadamente el 30% de esta variabilidad, sin embargo en el caso de algunos fármacos particulares este valor puede alcanzar incluso el 95% (3).

La farmacogenómica es el área encargada de estudiar cómo los factores genéticos interactúan entre sí mismos y con el ambiente para ocasionar que un individuo presente una respuesta distinta a otro ante un mismo tratamiento farmacológico. Por lo tanto, el desarrollo de esta área podría generar beneficios importantes para los sistemas de atención a la salud en el corto plazo a través de la selección del tratamiento más adecuado para cada paciente en particular, optimizando el uso de los recursos y permitiendo alcanzar la máxima eficacia terapéutica y reduciendo al mínimo la aparición de efectos adversos.

A través de diferentes enfoques, como son los estudios de asociación de genoma completo (GWAS por sus siglas en inglés) o estudios de genes candidato, se han identificado una amplia gama de variantes genéticas que determinan o pueden predecir las diferencias interpersonales en cuanto a efectividad terapéutica o toxicidad de los tratamientos contra diversos padecimientos. En cuanto a este aspecto cabe resaltar como hecho importante que algunas de estas variantes genéticas han demostrado poseer utilidad clínica para la estratificación de pacientes, donde con base en el genotipo de los individuos se predice la efectividad del tratamiento o se recomienda la dosis más adecuada para cada paciente específico (4).

Esto ha convertido a la farmacogenómica en una de las primeras aplicaciones donde se han trasladado con éxito los descubrimientos en el área genómica hacia el entorno de la práctica clínica diaria y el manejo individualizado de los pacientes. Sin embargo, como obstáculo para la adopción generalizada de la genotipificación de estas variantes se encuentra el hecho de que las variantes conocidas hasta el momento no lograr explicar completamente la gran variabilidad interpersonal en cuanto a respuesta ante los tratamientos. Una de las principales razones por lo que esto sucede es que la gran mayoría de las variantes fueron encontradas a través de estudios en poblaciones con ancestría europea (5) y dichas variantes no necesariamente se encuentran en la misma proporción en poblaciones con diferente ancestría u origen étnico y pueden variar en frecuencia aún en poblaciones muy similares (6). Además, se ha encontrado que en los genes que determinan la respuesta a fármacos existen variantes que tienden a encontrarse geográficamente localizadas y que pueden llegar a ser exclusivas de cada población (7).

Por esta razón para modelar el perfil farmacocinético de medicamentos de interés en individuos pertenecientes a la población mexicana resulta indispensable considerar no sólo las variantes genéticas conocidas, sino también aquellas que no han sido descubiertas aún al encontrarse únicamente en individuos pertenecientes a la población mexicana.

En este trabajo se demuestra que las variantes encontradas por secuenciación de nueva generación permiten generar modelos que representan adecuadamente fenómenos complejos tales como la variabilidad farmacocinética interindividual de medicamentos como la atorvastatina. Ésta es un fármaco importante en México debido a su uso en el tratamiento de hipercolesterolemia y la prevención de sus consecuencias cardiovasculares. Además, este trabajo realiza un primer paso para la generación de un catálogo de variantes genéticas presentes en 430 genes relacionados al metabolismo y respuesta a diversos fármacos y los efectos funcionales predichos y reportados que conllevan. La continuación y refinamiento de este primer catálogo de variantes será fundamental para optimizar en un futuro el uso de diversos fármacos en México a través del planteamiento de estrategias de terapia individualizada.

5. Marco teórico

5.1 Farmacocinética y farmacogenómica de atorvastatina

Las enfermedades cardiovasculares constituyen uno de los principales problemas de salud en México debido a su alta tasa de mortalidad (8), por lo que los medicamentos utilizados en su prevención y tratamiento forman parte de los fármacos más utilizados en nuestro país. La atorvastatina es uno de estos fármacos, indicado en el tratamiento de hipercolesterolemia ya que disminuye la progresión de ateroesclerosis y el consecuente riesgo cardiovascular mediante la reducción de los niveles de colesterol de baja densidad (LDL).

La atorvastatina pertenece a una familia de inhibidores de la 3-hidroxi-3-metilglutaril coenzima A (HMG-CoA) reductasa conocida con el nombre de estatinas. Esta familia de medicamentos actúa inhibiendo la síntesis de mevalonato a partir de HMG-CoA, el cual es el paso limitante de la síntesis endógena de colesterol. Aunque los miembros de esta familia de medicamentos son considerados generalmente seguros, hasta un 20% de los pacientes tratados con estatinas pueden presentar miopatía asociada al uso de estos fármacos, cuyos síntomas pueden ir desde ligera fatiga y dolor muscular hasta, en casos extremos, rabdomiolisis potencialmente fatal (9).

En el caso particular de atorvastatina, el desarrollo de miopatía ha sido relacionado con un perfil farmacocinético alterado de diversos metabolitos de este fármaco (10). Por lo tanto, anticipar el perfil farmacocinético esperado con base en el genotipo de un individuo ayudaría a realizar indicaciones terapéuticas (como un ajuste de dosis o un cambio a otro fármaco) que pudieran prevenir la aparición de estos efectos secundarios. Sin embargo, se ha mostrado en la población mexicana que un enfoque tradicional de farmacogenética donde se analice una variante de forma individual no es suficiente para explicar la amplia variabilidad en la farmacocinética de atorvastatina (11).

La atorvastatina es administrada por vía oral como una sal de calcio (forma abierta activa) en una dosis que oscila entre 10 y 80 mg/día. En condiciones fisiológicas, la forma abierta de la atorvastatina se encuentra en equilibrio con su forma lactona (inactiva) y ambas formas presentan un área bajo la curva de concentración en plasma a través del tiempo similares: 54-61 µg/L para la forma activa y 51-53 µg/L para la forma lactona (12).

Una vez administrada, el transporte y distribución de atorvastatina depende de diferentes productos génicos, entre los que se encuentran los transportadores de aniones orgánicos codificados en los genes *SLCO1B1*, *SLCO1B3* y *SLCO2B1*, los cuales presentan afinidad variable por las forma abierta y lactona de este fármaco (13,14). Con base en los niveles de expresión de estos genes en hepatocitos y considerando las actividades observadas *in vitro*, la contribución de estos tres transportadores a la absorción de atorvastatina y por lo tanto, su importancia para el metabolismo de este fármaco ha sido predicha como 52% para *SLCO1B1*, 42% para *SLCO1B3* y 6% para *SLCO2B1* (14).

En cuanto a metabolismo fase I –en el cual se modifican los grupos funcionales del fármaco principalmente por oxidación, reducción o hidrólisis (a diferencia del metabolismo fase II en el cual el fármaco se conjuga con otros compuestos)–, la atorvastatina es metabolizada a dos compuestos farmacológicamente activos (2-hidroxi-atorvastatina y 4-hidroxi-atorvastatina) y a sus lactonas correspondientes principalmente en los hepatocitos debido a la acción de la enzima codificada en el gen *CYP3A4* y en menor medida por la enzima codificada en *CYP3A5* (12,15) las cuales también presentan afinidades variables por las formas abierta y lactona de este fármaco.

Posteriormente, estos compuestos son glucuronidados por acción de diferentes UDP-glucuronosil transferasas, siendo las más importantes para el metabolismo de atorvastatina aquellas codificadas por los genes *UGT1A1*, *UGT1A3* y *UGT1A4*. Sin embargo, también se ha demostrado que la enzima codificada en *UGT2B7* tiene una actividad pequeña para la glucuronidación de este fármaco (16,17). Por su parte, la eliminación de atorvastatina se lleva a cabo principalmente a través de la bilis, debido a la acción de los transportadores glicoproteína-p y BCRP, codificados respectivamente en los genes *ABCB1* y *ABCG2* (12).

Por otro lado, en cuanto a la farmacogenómica de atorvastatina, se ha reportado que la presencia de polimorfismos en diferentes genes en su mayoría relacionados al transporte, distribución, metabolismo, respuesta y excreción de atorvastatina puede modificar un abanico de parámetros farmacológicamente importantes. Entre estos efectos se cuentan la tasa de reducción de lípidos (rs12003906 en *ABCA1* (18), la deleción rs1799752 en *ACE* (19), rs17238540 (20), rs12916 (21), rs3846662 (22), rs17671591 (23) en *HMGCR*, rs1057868 en *POR* (24), rs8175347 en *UGT1A3* (25), rs776746 en *CYP3A5* (26)), su perfil farmacocinético (rs2740574 en *CYP3A4* (27), rs4149056 (Val174Ala) en *SLCO1B1* (28), rs2231142 en *ABCG2* (29)) y la aparición de reacciones adversas (rs717620 en *ABCC2* (30), rs3892097 en *CYP2D6* (31)). Además, también existen reportes de polimorfismos asociados con otros efectos de atorvastatina, los cuales podrían afectar su farmacocinética a través de vías aún no dilucidadas completamente (por ejemplo las variantes rs1805094 en *LEPR* (32) y rs1800629 en *TNF* (33) están asociadas con un aumento en la densidad mineral ósea durante el tratamiento con atorvastatina).

De entre estas variantes que pueden modificar la absorción, distribución, metabolismo y excreción de las estatinas en general y de atorvastatina en particular cabe resaltar algunos ejemplos frecuentemente mencionados en la literatura científica como de gran importancia.

En primer lugar, la variante Val174Ala (con identificador rs4149056 en dbSNP) en el gen *SLCO1B1* ha sido asociada a la variación en el perfil farmacocinético de diferentes estatinas (34) y específicamente a la disminución en un 45% del aclaramiento de atorvastatina respecto a individuos homocigotos para el alelo de referencia (28) y el incremento de la concentración plasmática de atorvastatina (35). Asimismo, es de gran importancia el hecho que esta variante fue también asociada a la aparición de miopatía durante el tratamiento con simvastatina como resultado de un estudio de asociación de genoma completo en más de 12 mil pacientes (36).

Otra variante importante es la conocida como *CYP3A4*1B* (rs2740574), la cual se encuentra en la región promotora de este gen. El alelo G para esta variante está asociado con un incremento en la expresión de *CYP3A4* y un subsecuente aumento en el metabolismo de fase I debido a la monooxigenasa codificada en él (37). Del mismo modo, los individuos que presentan el alelo G para esta variante tienen menor probabilidad de requerir una disminución de dosis o un cambio de medicamento al utilizar atorvastatina y otras estatinas como simvastatina, lo cual se debe probablemente a este incremento en su metabolismo (27).

También, la variante *CYP3A5*3* (rs776746) afecta también el metabolismo fase I de las estatinas. Esta variante se localiza en el intrón 3 y crea un sitio críptico de “splicing”, causando la incorporación de parte de la secuencia de este intrón y un cambio de marco de lectura, el cual conlleva a su vez la aparición de un sitio de terminación prematuro y la pérdida de la actividad enzimática (38). Por esta razón, se ha propuesto que la presencia de esta variante afecta los niveles de respuesta en reducción de colesterol en individuos tratados con atorvastatina y otras estatinas metabolizadas por esta enzima. Sin embargo, se han encontrado resultados contradictorios, sugiriendo una asociación entre esta variante y una menor reducción de colesterol en pacientes hipercolesterolémicos brasileños (26) y el efecto contrario en un estudio anterior realizado en individuos caucásicos (39).

Por último, la variante Gln141Lys (rs2231142) en el gen *ABCG2* reduce la actividad de este transportador y por lo tanto afecta la farmacocinética de diversos fármacos (40), en el caso particular de atorvastatina esta variante se ha asociado a con un aumento en 72% del área bajo la curva de concentración plasmática a través del tiempo de este fármaco (29).

Debido a que la farmacocinética de atorvastatina puede ser modificada por el conjunto de variantes presentes en cualquiera de los genes involucrados en los diferentes pasos del metabolismo de atorvastatina, en este trabajo se utilizan las variantes genéticas presentes en los 20 genes mencionados anteriormente (*ABCB1*, *ABCC2*, *ABCG2*, *ACE*, *CYP2D6*, *CYP3A4*, *CYP3A5*, *HMGCR*, *LEPR*, *NOS3*, *POR*, *SLCO1B1*, *SLCO1B3*, *SLCO2B1*, *TNF*, *UGT1A1*, *UGT1A3*, *UGT1A4* y *UGT2B7*) que se sabe están involucrados en los efectos y metabolismo de atorvastatina, para modelar el perfil de la concentración plasmática de atorvastatina en 60 individuos mexicanos sanos, utilizando una regresión con vectores de soporte y una función de base radial como kernel.

6. ANTECEDENTES

En México, la investigación de los factores genéticos que afectan la variabilidad interindividual en el metabolismo y respuesta a fármacos se ha centrado, hasta ahora, en la determinación de la frecuencia con que se presentan variantes conocidas –y con efectos comprobados– en unos cuantos genes que incluyen los citocromos *CYP2D6*, *CYP2C9*, *CYP2C19*, *CYP3A4*, *CYP2E1* y las enzimas *NAT2* y *UGT1A4* (41).

Un factor a considerar es que el análisis de forma aislada de algunas variantes puede no lograr explicar la totalidad de las diferencias interpersonales en cuanto a efectividad o aparición de reacciones adversas, principalmente debido a que variantes adicionales en el mismo gen o en otros involucrados en diversas etapas del metabolismo, transporte y excreción de los fármacos pueden ejercer un efecto en la cinética, efectividad o toxicidad de un mismo medicamento.

Las técnicas de secuenciación constituyen el máximo nivel de caracterización de los ácidos nucleicos, ya que posibilitan la detección de mutaciones puntuales, inserciones, deleciones y translocaciones aún sin conocimiento previo de su existencia. Por esta razón la secuenciación ha demostrado ser una estrategia útil para analizar la diversidad genética en genes importantes en el metabolismo de fármacos, permitiendo simultáneamente el análisis de la distribución de las variantes conocidas y el descubrimiento de variantes novedosas con posible utilidad clínica (42).

A pesar de estas ventajas, la secuenciación de los genes con importancia en farmacogenómica en grupos de individuos pertenecientes a la población mexicana o estadounidenses con ascendencia mexicana ha sido hasta el momento bastante limitada y se ha emprendido por separado para sólo un pequeño número de genes que intervienen en el metabolismo y respuesta de fármacos específicos (41).

En estos estudios previos, aun cuando el análisis se ha limitado a un par de genes, la secuenciación ha permitido la identificación de variantes genéticas no reportadas previamente y que se presentan únicamente en individuos con ascendencia mexicana para los genes *DCK* y *CMPK* que intervienen en la activación del fármaco antineoplásico gemcitabine (43) y en el gen *MTHFR* que puede afectar la sensibilidad a fármacos agonistas del ácido fólico, como por ejemplo metotrexato (44).

También, a través de la secuenciación, ha sido posible identificar variantes no sólo novedosas, sino con posibles efectos predichos en el metabolismo de antidepresivos, antipsicóticos, antihipertensivos y otros fármacos metabolizados por la proteína codificada en el gen *CYP2D6* (42). Del mismo modo, la secuenciación ha permitido el descubrimiento de variantes con asociación a una mejor respuesta a terapia antidepresiva con fluoxetina o desipramina a través de la secuenciación del gen *BDNF* (45) y de los genes *ABCB1*, *SLC6A2*, *SLC6A3*, *SLC6A4*, *CREB1*, *CRHR1* y *NTKR2* (46), así como también algunas variantes que reducen la absorción *in vitro* de compuestos cuyo transporte depende de la proteína codificada en *SLC47A1* como el herbicida paraquat y el fármaco antidiabético metformina (47).

El desarrollo de las técnicas de secuenciación de ácidos nucleicos y particularmente el surgimiento de las tecnologías de “nueva generación” (“Next-Generation Sequencing”), las cuales se distinguen por una capacidad masiva de producción de información mediante la secuenciación en paralelo de millones de moléculas de ADN, ha resultado en un abaratamiento dramático de los costos por base secuenciada (48), permitiendo abordar diseños experimentales que abarcan cientos o miles de genes, o incluso el genoma completo de los organismos.

Con la intención de aprovechar al máximo la capacidad masiva de secuenciación que ofrecen las tecnologías de “nueva generación” han surgido métodos adicionales que reducen la complejidad de las muestras para secuenciar sólo un subconjunto de los ácidos nucleicos presentes. Estas variantes metodológicas consisten en sistemas de enriquecimiento por hibridación de sondas complementarias o amplifiación diferencial que “capturan” sólo los fragmentos de ADN que se desea secuenciar.

De forma análoga a lo que se conoce como secuenciación del “exoma” –donde se lleva a cabo un análisis de todas las regiones codificantes para proteína (49)–, es posible utilizar estos sistemas de enriquecimiento para “capturar” todos los genes relevantes para el metabolismo de fármacos o incluso concentrarse solamente en aquellas regiones donde el efecto funcional de las posibles variantes sea más sencillo de predecir y conlleve mayor significado biológico, por ejemplo, los sitios de “splicing” o algunas regiones no transcritas importantes como sitios de unión a factores de transcripción (50).

En este trabajo se utilizan tecnologías de nueva generación para la secuenciación comprensiva, por primera vez en individuos mexicanos, de las regiones con mayor significado biológico de 430 genes que intervienen en el metabolismo o mecanismo de acción de un gran número de fármacos, permitiendo modelar la farmacocinética de atorvastatina en 60 individuos y caracterizar la diversidad genética en cuanto al análisis de la presencia de variantes previamente reportadas y la identificación de nuevas variantes importantes para el desarrollo de la farmacogenómica en esta población.

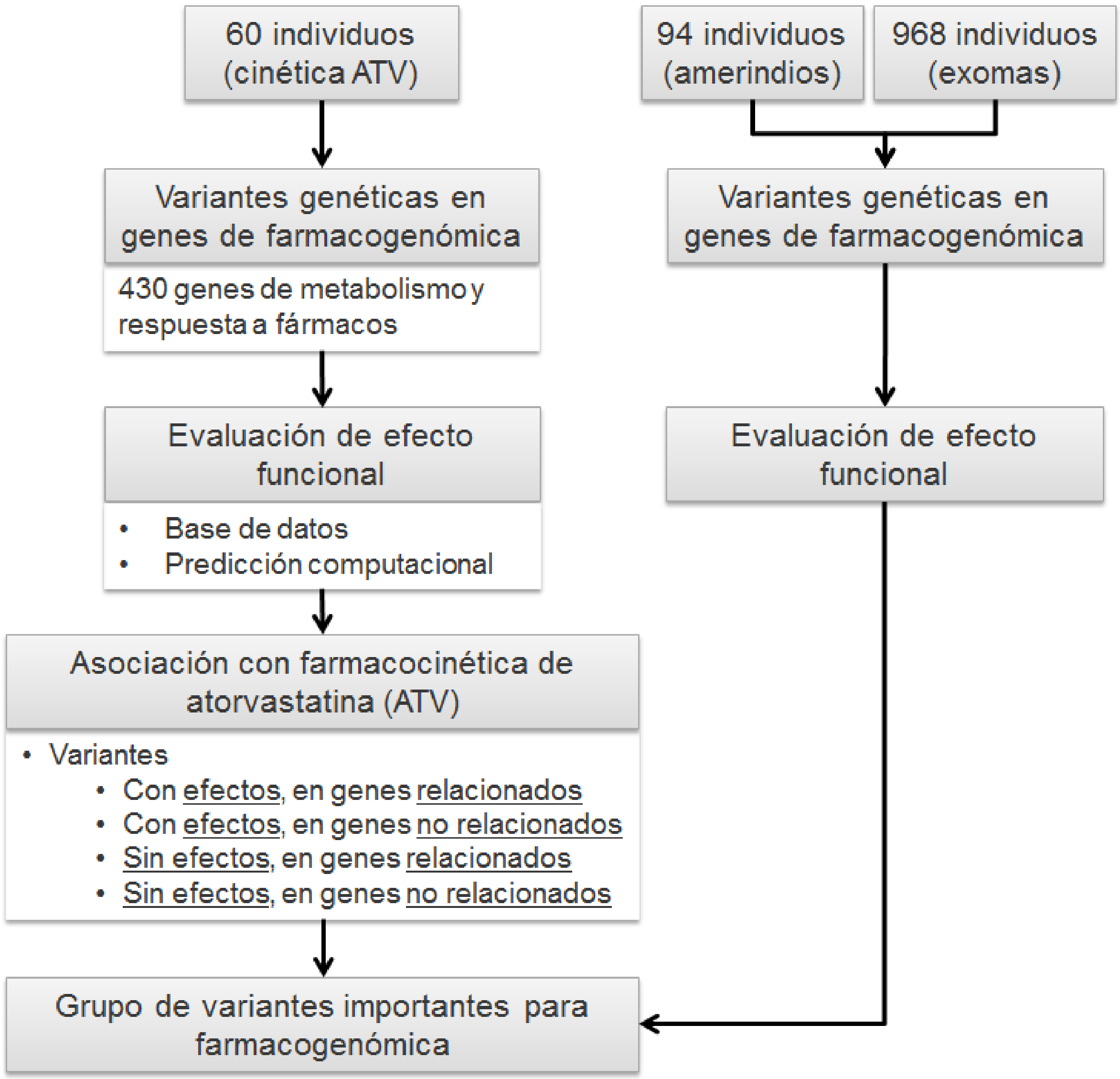
7. METODOLOGÍA

7.1 PANORAMA GENERAL

Este trabajo tiene como objetivo en primer lugar, demostrar la utilidad de la colección de diversidad genética en genes involucrados en el metabolismo y respuesta de diversos fármacos en muestras de población mexicana mediante el modelado de la farmacocinética de atorvastatina en 60 individuos. Para ello se diseñó un panel de enriquecimiento para secuenciar las regiones correspondientes a un grupo de 430 genes en un grupo de individuos caracterizados farmacocinéticamente para atorvastatina. Los posibles efectos de las variantes encontradas fueron evaluados a través de los reportes de asociación a fármacos en la base de datos de PharmGKB (efectos reportados) y mediante una combinación de herramientas computacionales (efectos predichos). Las variantes con efectos funcionales predichos o reportados que se encuentran en genes relacionados al metabolismo y respuesta ante atorvastatina fueron utilizadas para construir un modelo de regresión con vectores de soporte que pudiera representar adecuadamente la variabilidad farmacocinética de este fármaco en los 60 individuos. El desempeño de este modelo de regresión fue comparado contra el obtenido por modelos generados a partir de diferentes grupos de variantes, demostrando la utilidad de la evaluación de los efectos funcionales de las variantes encontradas por secuenciación para este fin.

Además, con este proyecto se comienza la recopilación de un catálogo de las variantes genéticas presentes en 430 genes relacionados a medicamentos para informar subsecuentes estudios de farmacogenómica en mexicanos. Para complementar este catálogo se añadieron las variantes encontradas en las mismas regiones de un grupo de 94 genomas completos de individuos amerindios y 968 exomas provenientes de otros proyectos de secuenciación. La evaluación de los efectos funcionales de cada una de estas variantes se realizó también de la manera antes descrita y se evaluó la frecuencia promedio a nivel mundial a través de la base de datos de kaviar. El catálogo de variantes con efectos funcionales predichos o reportados representa un listado preliminar de las variantes importantes para farmacogenómica en la población mexicana ya que podrían modificar la farmacocinética, efectividad o toxicidad de fármacos.

Figura 1. Diagrama de flujo del proyecto



7.2 Sujetos

Se incluyeron en este estudio 60 voluntarios aparentemente sanos del noreste de México: hombres de 18 a 45 años de edad, seronegativos para VIH, VHB y VHC, con nivel de peso normal (IMC entre 20 y 25 kg/m^2), con resultados normales de biometría hemática, química sanguínea y análisis general de orina, participantes en un estudio de la farmacocinética de atorvastatina (Australian New Zealand Clinical Trials Registry ACTRN12614000851662). El protocolo de investigación fue aprobado por el comité de investigación y ética del Instituto Nacional de Medicina Genómica (INMEGEN, México). El ADN fue extraído de sangre y fue sometido a enriquecimiento y secuenciación de nueva generación como se describe más adelante.

Para complementar la caracterización de la diversidad genética se analizaron además datos provenientes de otros proyectos de secuenciación llevados a cabo en el INMEGEN. De estos proyectos se incluyeron las variantes encontradas en 968 exomas completos de individuos mexicanos secuenciados como parte del Consorcio SIGMA de Diabetes Tipo 2 (51) y 94 genomas completos de individuos con ancestría amerindia (residentes originarios de una comunidad indígena que hablan la lengua correspondiente y cuyos padres y abuelos también son originarios de dicha comunidad, datos aún no publicados).

7.3 Perfil farmacocinético de atorvastatina

La obtención del perfil farmacocinético para atorvastatina en los 60 voluntarios ha sido descrita previamente (11). En resumen, a los voluntarios se les administró una dosis única de 80 mg de atorvastatina después de una noche de ayuno. Se tomaron muestras de sangre antes de la administración del fármaco y a las 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0, 8.0, 12.0, 24.0, 36.0 y 48.0 horas. Estas muestras de sangre fueron mezcladas con 4 volúmenes de acetonitrilo y centrifugadas para separar el plasma. El sobrenadante fue recuperado y sometido a análisis en un instrumento Agilent 1100 de cromatografía de alta eficiencia (HPLC) equipado con una bomba binaria y automuestreador y conectado a un instrumento Agilent 6410 de espectrometría de masas en tándem (MS/MS) con un sistema de detección con triple cuadrupolo para medir los niveles de atorvastatina en su forma abierta. El área bajo la curva de concentración de atorvastatina desde el tiempo 0 al tiempo de la última determinación (AUC) fue estimada considerando un método no compartimental y utilizando el método lineal-logarítmico trapezoidal para el cálculo.

7.4 Preparación de bibliotecas, enriquecimiento y secuenciación de nueva generación

Un grupo de 430 genes fue seleccionado de tal forma que se incluyeran aquellos que codifican enzimas que intervienen en la absorción, distribución, metabolismo y excreción de diferentes fármacos (tales como enzimas metabolizadoras de fase I y fase II, receptores y transportadores) así como sus blancos terapéuticos. Además, se incluyeron aquellos genes presentes en dos plataformas comerciales de genotipificación farmacogenómica utilizadas en la actualidad: el microarreglo DMET de la compañía Affymetrix (52) y el microarreglo PharmaChip de la compañía Progenika, así como los reportados como importantes en la principal base de datos especializada en farmacogenómica: PharmGKB (53).

Después se diseñó un panel de sondas de enriquecimiento Haloplex a la medida (Agilent Technologies, Santa Clara, CA) que incluye las regiones exónicas, los sitios de “splicing”, las regiones no transcritas 5’ y 3’ y aquellas donde se presentan variantes previamente reportadas del grupo de 430 genes seleccionado por su intervención en el metabolismo y respuesta ante diversos medicamentos y un panel de 446 SNPs que permiten estimar la contribución de las poblaciones amerindia, europea y africana a la ancestría de las poblaciones latinoamericanas actuales (54) para poder evaluar si la ancestría se encuentra asociada con la farmacocinética de atorvastatina.

Utilizando este panel de enriquecimiento se prepararon bibliotecas de secuenciación de nueva generación para los 60 individuos caracterizados farmacocinéticamente de acuerdo al protocolo del fabricante. En resumen, 225 ng de ADN genómico extraído desde sangre total fueron en un primer paso fragmentado con ocho mezclas de enzimas de restricción diferentes y después sometidos a hibridación con oligonucléotidos "barcode" que permiten la identificación de la muestra y con el panel de sondas biotiniladas HaloPlex para circularizar los fragmentos que abarcan las regiones objetivo. Estos fragmentos circularizados son capturados mediante perlas magnéticas recubiertas de estreptavidina y, posteriormente, amplificados por PCR para incluir dentro de los productos las secuencias de los oligonucleótidos "barcode" y los adaptadores de secuenciación. La concentración de las librerías fue verificada en un instrumento Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). La secuenciación de las bibliotecas se realizó en un equipo Genome Analyzer IIx (Illumina Inc., San Diego CA) con una longitud de lectura de 150 pb en formato pareado ("paired-end").

7.5 Análisis de secuencia

7.5.1 Llamado de variantes

En un primer paso se analizó la calidad de la corrida de secuenciación de acuerdo a la longitud de las lecturas obtenidas y la calidad de la asignación de bases por posición a través de las lecturas. Después, las lecturas fueron recortadas para remover regiones de baja calidad y las secuencias pertenecientes a los adaptadores de secuenciación mediante el programa Trimmomatic (55). La eficiencia del enriquecimiento en las regiones genómicas deseadas fue evaluada a través de las métricas obtenidas con el programa CalculateHsMetrics dentro del paquete de software Picard (picard-tools-1.130, https://github.com/broadinstitute/picard). Posteriormente se realizó un llamado de variantes de acuerdo con los protocolos de "buenas prácticas" del paquete GATK con la única diferencia siendo que no se marcaron lecturas como duplicadas debido a que se esperaban posiciones específicas de inicio y final de lecturas a causa de la fragmentación enzimática llevada a cabo durante la preparación de bibliotecas. En resumen, las lecturas que pasaron los filtros de calidad fueron alineadas contra el genoma humano de referencia (versión hg19) con el programa BWA-mem (http://arxiv.org/abs/1303.3997), después se realizó un realineamiento alrededor de las regiones con inserciones y deleciones, el posterior recalibrado de calidad de bases, un primer llamado de variantes de un único nucleótido, inserciones y deleciones, la genotipificación conjunta de todos los individuos y el recalibrado del puntaje de calidad de variantes mediante las herramientas incluidas en la paquetería de GATK (56) versión 3.3-0-g37228af, utilizando los parámetros recomendados (57).

7.5.2 Ancestría

Para estimar las proporciones de los componentes ancestrales en los individuos secuenciados en este estudio se utilizó el programa admixture (58), considerando únicamente los 446 polimorfismos de un solo nucleótido (SNPs) incluidos en el panel mencionado anteriormente (54) que permite estimar la contribución de las poblaciones amerindia, europea y africana a la ancestría de las poblaciones latinoamericanas actuales.

Como individuos de referencia para este análisis se seleccionaron de forma aleatoria individuos pertenecientes a diferentes poblaciones del proyecto de los 1,000 genomas (59) (2 individuos residentes de Utah, EEUU con ascendencia de Europa del Norte u Oeste (CEU), 2 individuos Toscanos de Italia (TSI) y 2 individuos Ibéricos de España (IBS) como referencia del componente europeo, 2 individuos Yoruba de Nigeria (YRI), 2 individuos Luhya de Kenya (LWK) como referencia del componente africano) y un grupo de 6 individuos con ascendencia amerindia (1 Tepehuana, 1 Maya, 1 Nahua, 1 Tarahumara, 1 Totonaca y 1 Zapoteca) reclutados por el INMEGEN (60). Además, para fines de comparación se seleccionaron aleatoriamente 2 individuos pertenecientes al proyecto de 1,000 genomas (59) con ascendencia mexicana de Los Angeles (MXL) como controles positivos de ancestría.

7.5.3 Predicción de efecto funcional

Se consideró que las variantes encontradas podrían tener un efecto funcional si se encontraban reportadas en la base de datos de PharmGKB (53) como asociadas al metabolismo, respuesta o toxicidad de algún fármaco o si tenían una alta probabilidad de presentar un efecto funcional de acuerdo a una combinación de herramientas computacionales: una clasificación como deletérea según el algoritmo de PROVEAN (61), un método basado en la conservación de aminoácidos en un alineamiento múltiple de secuencias homólogas que permite analizar el efecto de las variantes codificantes y cuya principal ventaja sobre otros algoritmos como SIFT (62) y PolyPhen (63) es la capacidad de analizar el efecto funcional no sólo de substituciones de aminoácidos, sino también las inserciones y deleciones de los mismos; que conllevan un impacto “alto” según el algoritmo de Variant Effect Predictor (VEP) de ENSEMBLE (64), un método que mapea las variantes a los transcritos conocidos y les asigna un nivel de impacto de acuerdo al efecto que tienen sobre los mismos y un puntaje mayor a 0.5 según FATHMM-MKL (65), un método de aprendizaje automatizado que utiliza una combinación de los grupos de datos de ENCODE para asignar una probabilidad de que la variante tenga un efecto deletéreo sobre elementos codificantes y no codificantes. Estas herramientas de predicción de efectos funcionales (PROVEAN, VEP y FATHMM-MKL) fueron utilizadas con los parámetros por defecto recomendados para cada una de ellas.

7.6 Caracterización de la diversidad genética en 430 genes importantes de farmacogenómica

Como se mencionó anteriormente, para complementar la caracterización de la diversidad genética a los 60 individuos con perfil farmacocinético de atorvastatina se agregaron datos provenientes de 968 exomas completos de individuos mexicanos secuenciados como parte del Consorcio SIGMA de Diabetes Tipo 2 (51) y 94 genomas completos de individuos con ancestría amerindia (datos aún no publicados). De estos datos se retuvieron solamente las variantes genéticas presentes en las regiones correspondientes al panel de enriquecimiento descrito en las secciones anteriores y se evaluó la frecuencia de las variantes genéticas en cada uno de ellos.

Para obtener la frecuencia de las variantes genéticas a nivel global, se utilizó la base de datos kaviar, la cual incluye información de más de 13 mil genomas completos y 64 mil exomas provenientes de todas partes del mundo (66). Esta información se contrastó, después, con la frecuencia con la cual se presentan las variantes genéticas en individuos amerindios.

7.7 Asociación de variantes genéticas con la farmacocinética atorvastatina

7.7.1 Construcción del modelo de regresión utilizando máquinas de vectores de soporte

Se seleccionaron las variantes presentes en 20 genes relacionados con la respuesta terapéutica, el metabolismo u otros efectos de atorvastatina (*ABCA1*, *ABCB1*, *ABCC2*, *ABCG2*, *ACE*, *CYP2D6*, *CYP3A4*, *CYP3A5*, *HMGCR*, *LEPR*, *NOS3*, *POR*, *SLCO1B1*, *SLCO1B3*, *SLCO2B1*, *TNF*, *UGT1A1*, *UGT1A3*, *UGT1A4* y *UGT2B7*) que tienen un efecto funcional predicho por alguna de las herramientas utilizadas (PROVEAN, VEP o FATHMM-MKL) o que están reportadas en PharmGKB como asociadas al metabolismo, respuesta o toxicidad de algún fármaco para construir un modelo de regresión con vectores de soporte usando una función de base radial (RBF) como kernel para modelar el área bajo la curva de concentración plasmática (AUC) de atorvastatina.

El método de vectores de soporte es una herramienta universal para resolver problemas de estimación de funciones no lineales. En este tipo de algoritmos los datos, que se encuentran relacionados de manera no lineal a la variable predictora en el espacio original, son mapeados a través de una función de mapeo (conocida como función kernel) a un espacio con dimensionalidad superior donde puede construirse una regresión multivariada lineal (67).

La función de regresión resultante se desvía de todos los datos de entrenamiento como máximo un valor de tolerancia determinado, sin embargo, sólo los datos con desviaciones igual al límite de la tolerancia (conocidos como vectores de soporte) contribuyen al modelo. Este tipo de algoritmos son especialmente útiles para el manejo de datos con un gran número de dimensiones, ya que el modelo no depende del número original de dimensiones, sino solamente de un pequeño subconjunto de los datos (los vectores de soporte) (68).

Los vectores de soporte son ideales para el análisis de sistemas biológicos debido a que son capaces de resolver problemas de reconocimiento de patrones involucrando tamaños de muestra reducidos, relaciones no lineales entre las variables y datos originales con un gran número de dimensiones (69).

Los valores de AUC fueron transformados con logaritmo natural y utilizados como valores objetivo. El número de alelos alternativos para cada una de las variantes –presentes en genes relacionados con atorvastatina y con efectos funcionales predichos o reportados– fueron utilizados para construir el modelo de regresión con vectores de soporte utilizando una función de base radial como kernel.

Los hiperparámetros del modelo: C (el cual controla las concesiones mutuas entre el valor hasta el cual se permiten las transgresiones a la tolerancia y la complejidad del modelo) y gamma (que controla magnitud de la influencia que cada uno de los datos tiene sobre el modelo) deben optimizarse para evitar un sobreajuste. Para la determinación de los valores de los hiperparámetros durante la construcción del modelo se puede obtener un estimado del desempeño del modelo mediante una validación cruzada de k pliegues. En este procedimiento, en lugar de dividir el grupo de datos iniciales en grupos separados de datos (datos de entrenamiento y de prueba) lo que reduciría drásticamente el número de muestras utilizadas en el entrenamiento, el grupo de datos iniciales se divide en k grupos más pequeños. Después se deja uno de estos grupos fuera para su uso como set de prueba y se entrena el modelo con los grupos restantes, iterando hasta que todos los grupos han sido utilizados como set de prueba una vez y la métrica del desempeño del modelo es el promedio de todos los valores calculados. Al utilizar este procedimiento es posible optimizar el valor de los hiperparámetros del modelo de tal forma que se favorece la más alta reproducibilidad en nuevos datos aún si el número inicial de muestras es relativamente limitado.

Los hiperparámetros del modelo C y gamma fueron seleccionados utilizando una estrategia de rejilla de búsqueda. En resumen, para cada hiperparámetro se seleccionaron nueve valores equidistantes entre 10^{-4} y 10^{4} y se construyó un modelo con cada combinación de valores considerando una validación cruzada de 3 pliegues, seleccionando como valores finales una combinación con valores adecuados de coeficiente de correlación. Adicionalmente, continuamos la optimización del parámetro gamma al graficar los valores de error cuadrático medio para los sets de entrenamiento y prueba, seleccionando un valor de gamma que resultara en errores cuadráticos medios relativamente bajos para ambos sets y mayor reproducibilidad en el set de prueba (a través de una menor desviación estándar en el error cuadrático medio de dicho set).

Como siguiente paso, calculamos la importancia de cada variante dentro del modelo original con la intención de evaluar si es necesario contar con todas las variantes con efectos funcionales presentes en los genes relacionados a atorvastatina o si un subgrupo más pequeño sería suficiente para conseguir un buen modelo de la farmacocinética de atorvastatina. Con este objetivo construimos modelos de regresión adicionales dejando fuera una de las variantes por turno y calculamos la pérdida en varianza explicada de estos modelos reducidos respecto al modelo original completo. Usando esta información refinamos el modelo de regresión al seleccionar únicamente las variantes más importantes y evaluando cuántas variantes son necesarias para explicar una proporción adecuada de la varianza.

Para evaluar la validez interna del modelo de regresión refinado, seleccionamos de forma aleatoria un número de variantes (con efectos funcionales presentes en los genes relacionados a atorvastatina) igual al incluido en el modelo refinado y construimos un modelo de regresión utilizando las variantes aleatorias, iterando hasta 10^7 modelos de hipótesis nula para calcular el valor p asociado a la obtención de un modelo con varianza explicada igual o mayor que aquella obtenida por el modelo refinado final.

Finalmente, para contrastar el desempeño del modelo final construido a partir de las variantes con efectos funcionales predichos o reportados que se encuentran en 20 genes relacionados con el metabolismo y respuesta de atorvastatina contra el que se obtendría al considerar otros subgrupos de variantes, realizamos nuevamente todos los pasos para la construcción y refinación de un modelo de regresión con vectores de soporte a partir de tres grupos adicionales de variantes genéticas: 1) variantes que se localizan en 20 genes relacionados con atorvastatina que no presentan efectos funcionales predichos o reportados; 2) variantes con efectos funcionales en genes no relacionados con atorvastatina y 3) variantes sin efectos funcionales presentes en genes no relacionados con atorvastatina. La construcción de todos los modelos de regresión con vectores de soporte fueron realizados en Python mediante el uso de las herramientas de aprendizaje automatizado de scikit-learn (70).

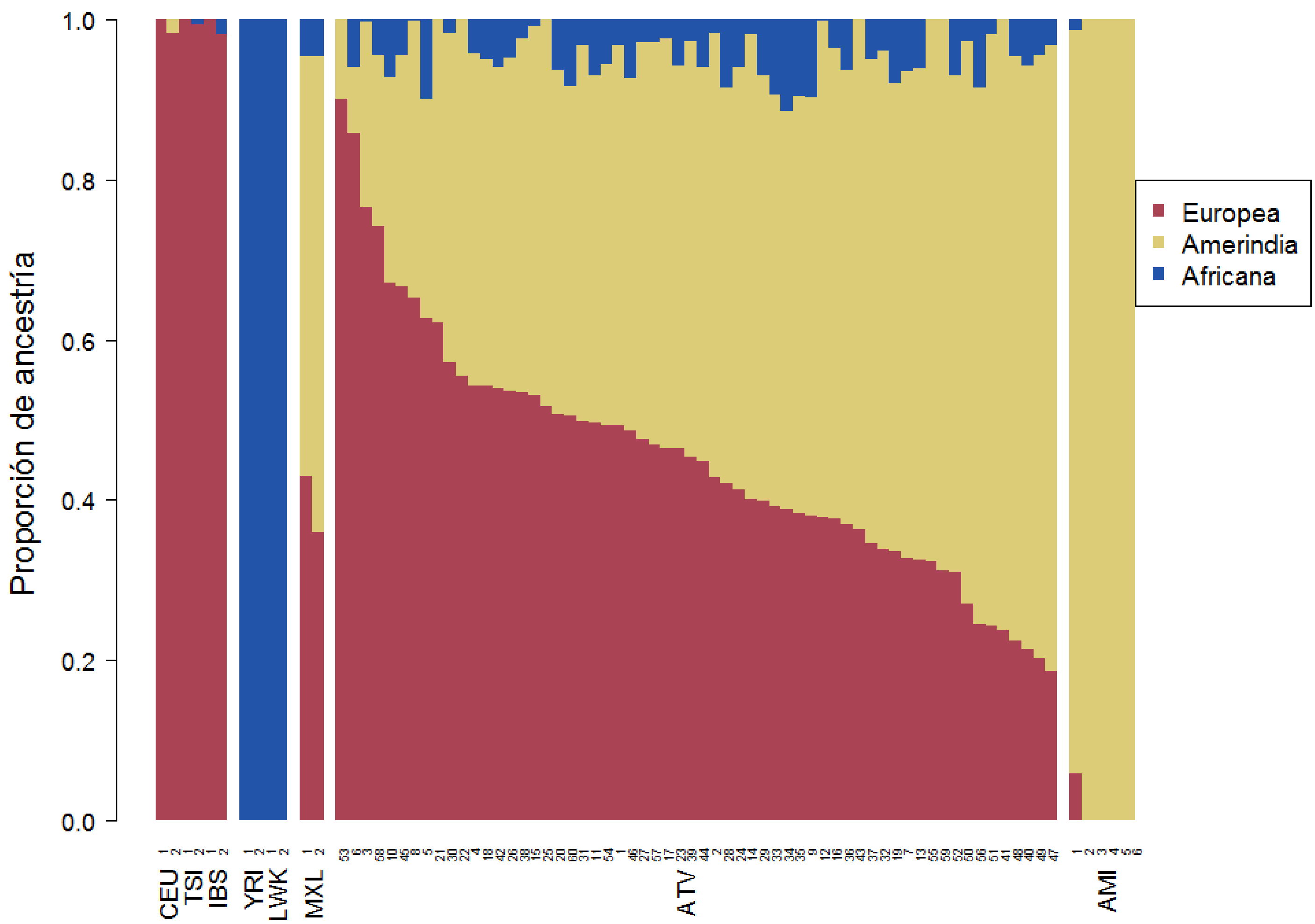
8. Resultados

En esta sección se muestran en idioma español resultados publicados en el artículo derivado del trabajo de esta investigación (71) (Anexo 1), incluyendo también resultados adicionales no publicados en dicho artículo.

8.1 Distribución de componentes ancestrales de los 60 individuos caracterizados farmacocinéticamente

Los componentes ancestrales de los 60 individuos caracterizados para la farmacocinética de atorvastatina presentaron una mediana de 49.90% (RI: 42.62%-60.75%) para el componente amerindio, de 45.15% (RI: 34.44%-53.48%) para el componente europeo y de 4.52% (RI: 1.97%-6.48%) para el componente africano.

Figura 2. Componentes ancestrales de los individuos caracterizados farmacocinéticamente para ATV



De izquierda a derecha, contribución de los componentes europeo, amerindio y africano a la ancestría de los individuos provenientes del proyecto de los 1000 genomas utilizados como referencia del componente europeo (2 CEU, 2 TSI, 2 IBS), referencia del componente africano (2 YRI, 2 LWK), controles de ancestría mexicana (2 MXL), individuos caracterizados farmacocinéticamente para atorvastatina (60 ATV) y 6 individuos amerindios utilizados como referencia del componente amerindio (AMI).

Las proporciones de ancestría europea, amerindia y africana a nivel individual fueron calculadas con la esperanza de encontrar una asociación con la farmacocinética de atorvastatina. Sin embargo, estos valores no se encuentran relacionadas linealmente con los valores de AUC de atorvastatina (R^2<0.002) ni provocan una modificación importante de los modelos de regresión con vectores de soporte generados en pasos posteriores.

8.2 Asociación al perfil farmacocinético de atorvastatina

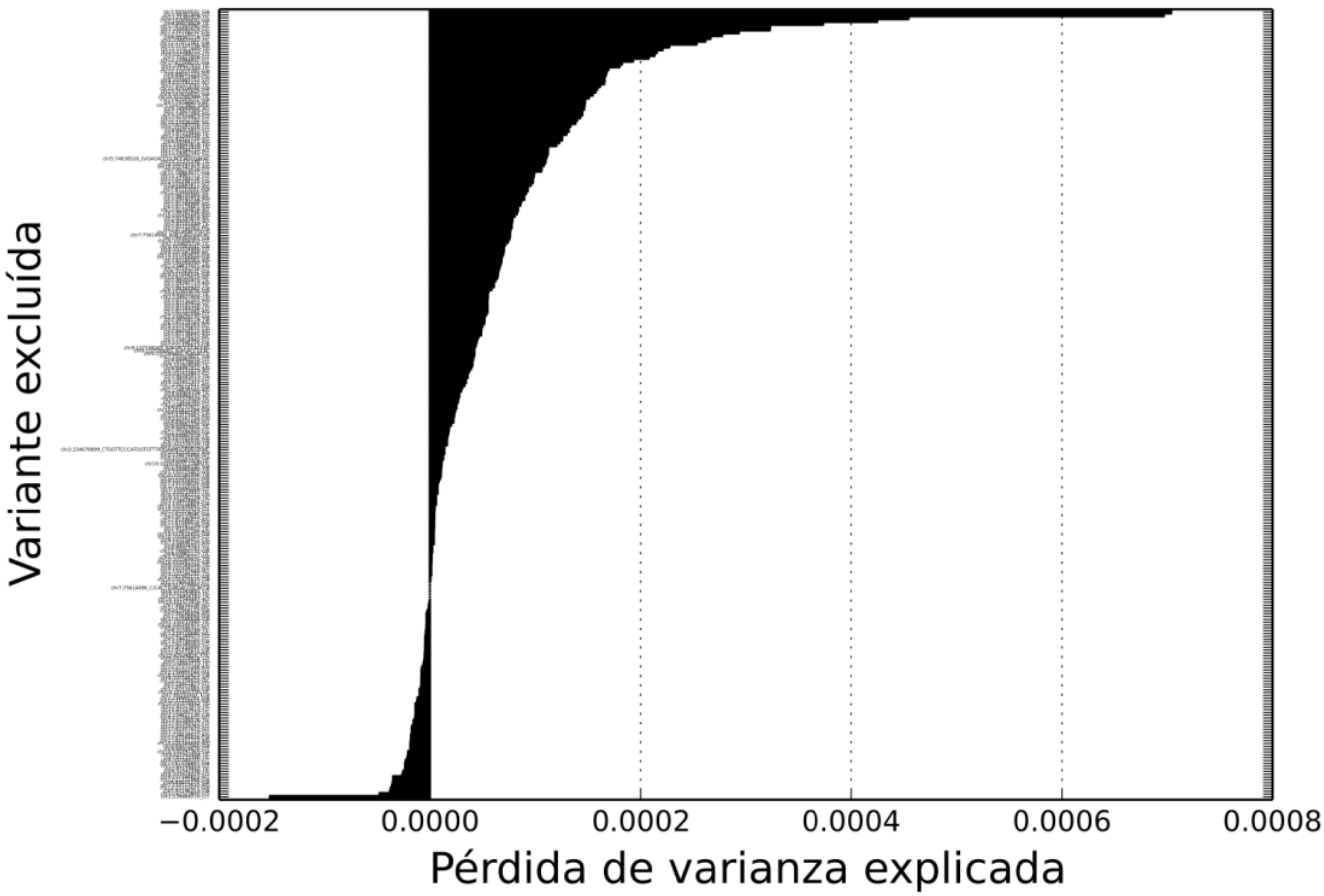
8.2.1 Construcción y refinación del modelo de regresión

En los 20 genes relacionados directamente con el metabolismo y respuesta de atorvastatina se encontró un total de 1,161 variantes en los 60 individuos, de las cuales el 24.03% (279 variantes) tienen efectos funcionales predichos o reportados. De forma importante, el 13.26% (37 variantes) de estas variantes con efectos funcionales no se encuentran reportadas en dbSNP (versión 144) y pueden considerarse como nuevas.

Como se mencionó con anterioridad, en primer lugar se construyó un modelo original de regresión con máquinas de vectores de soporte construido utilizando el logaritmo de AUC como valores objetivos y el número de alelos alternativos presentes para cada una de estas 279 variantes con efectos funcionales presentes en genes relacionados con atorvastatina. Este modelo de regresión con máquinas de vectores de soporte utiliza una función de base radial (RBF) como kernel e hiperparámetros C=1 y gamma=$1x10^{-1.3}$ y permite explicar el 95.96% de la varianza.

Como siguiente paso, construimos 279 modelos de regresión adicionales, dejando fuera del modelo la información de una variante diferente en cada uno de ellos. Después calculamos la pérdida de varianza explicada de estos modelos reducidos respecto al modelo original completo, permitiendo de esta manera evaluar la importancia la importancia de cada una de las variantes en el modelo original.

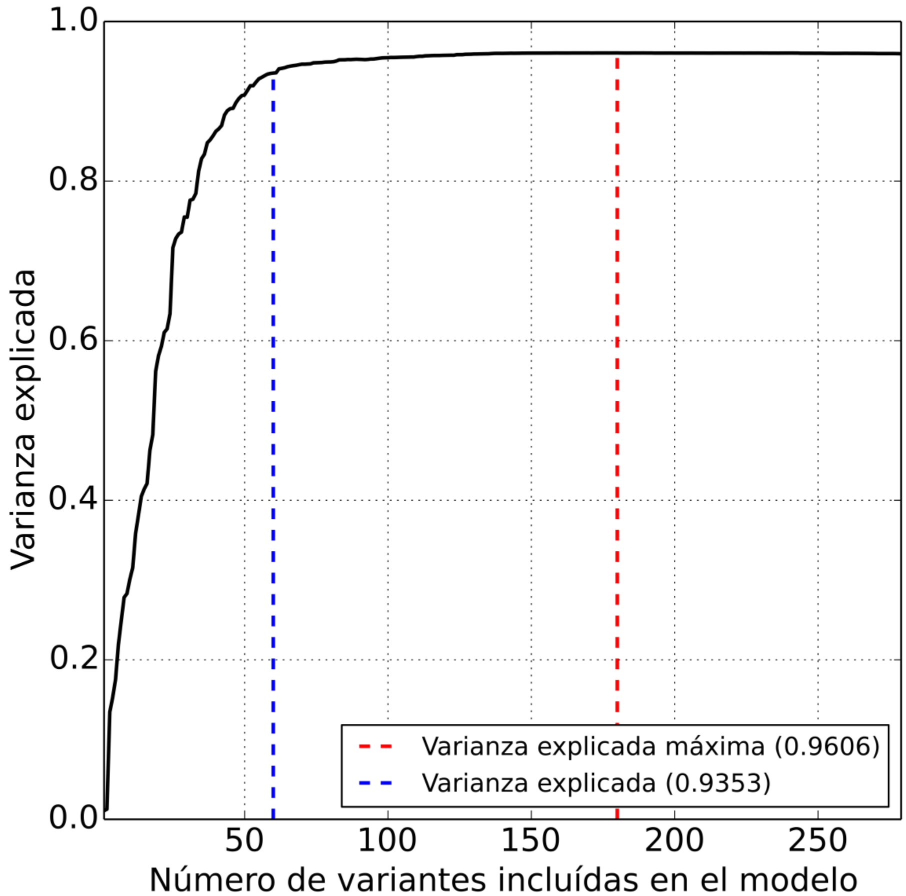
Figura 3. Pérdida de varianza explicada de los modelos reducidos respecto al modelo original completo como indicador de la importancia de cada variante



Pérdida de varianza explicada de cada uno de los modelos reducidos respecto al modelo original completo. La mayor pérdida de varianza explicada indica la mayor importancia en el modelo completo.

Después seleccionamos un número específico de variantes en orden de importancia hasta llegar al total, para evaluar el número de variantes con el cual se explica una proporción importante de la varianza. Este análisis mostró que al construir un modelo utilizando las 180 variantes más importantes se logra explicar la máxima proporción de la varianza (96.06%), la cual es incluso más alta que aquella explicada por el modelo original completo que utiliza todas las variantes (95.96%). Además, este análisis también mostró que es posible reducir el número de variantes incluidas en el modelo a sólo un tercio sin causar una reducción drástica en el valor de la varianza explicada, por lo que se consideraron únicamente las primeras 60 variantes para construir el modelo final refinado que logra explicar el 93.53% de la varianza.

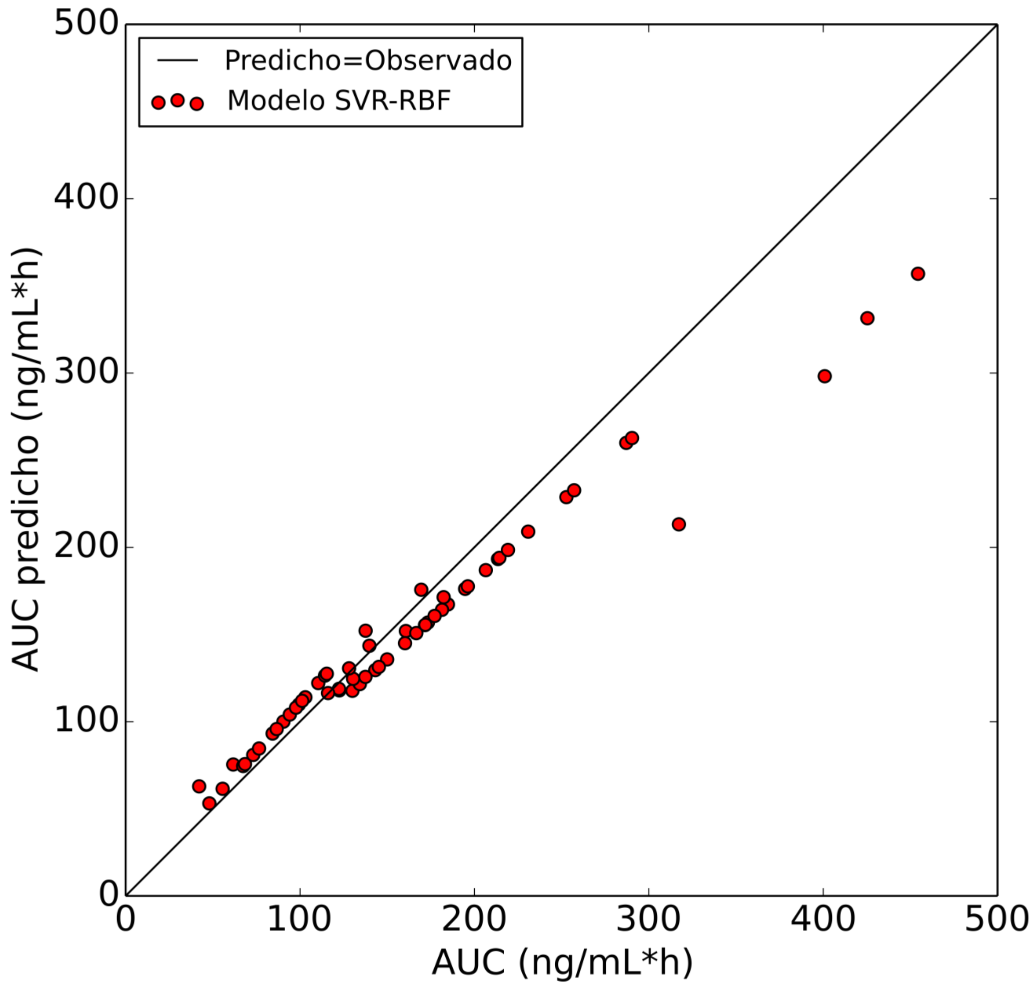
Figura 4. Varianza explicada por modelos construidos al agregar las variantes con mayor importancia hasta alcanzar el total



Varianza explicada por modelos construidos al agregar una a una las variantes con mayor importancia hasta alcanzar el total (279 variantes). Notablemente, la varianza explicada por el modelo construido con las primeras 60 variantes es de 93.53%.

Como se muestra en la figura siguiente, este modelo de regresión refinado que considera únicamente las 60 variantes más importantes muestra una buena correlación con los datos de AUC observados experimentalmente en los 60 individuos (R^2=0.884).

Figura 5. Comparación entre los valores de AUC predichos por el modelo final de regresión y aquellos obtenidos experimentalmente para cada uno de los individuos



Valores de AUC predichos por el modelo final de regresión y valores obtenidos experimentalmente para cada uno de los individuos. La línea negra representa una correlación perfecta entre valores predichos y observados.

Cabe resaltar que de las 60 variantes incluidas en el modelo final refinado, el 38.33% (23 variantes) no se encuentran reportadas como asociadas a ningún fármaco en la base de datos de PharmGKB y cuentan sólo con efectos funcionales predichos por alguna de las herramientas de predicción utilizadas. Además, solamente una (1.67%) de estas 60 variantes es nueva aunque en el modelo se encuentran incluidas 5 variantes (8.33%) que pueden considerarse raras en esta población al presentar una frecuencia menor al 5%.

Por otro lado, ninguna de las variantes incluidas en el modelo de regresión presentó una asociación significativa ($p<0.05$) con los valores de AUC según una prueba de Kruskal-Wallis a nivel de genotipo y al realizar un análisis individual haciendo referencia a los pasos de la vía de procesamiento del fármaco que pudieran estar afectando las variantes genéticas no se llegó a ningún resultado concluyente (datos no mostrados).

8.2.2 Validación interna del modelo de regresión refinado

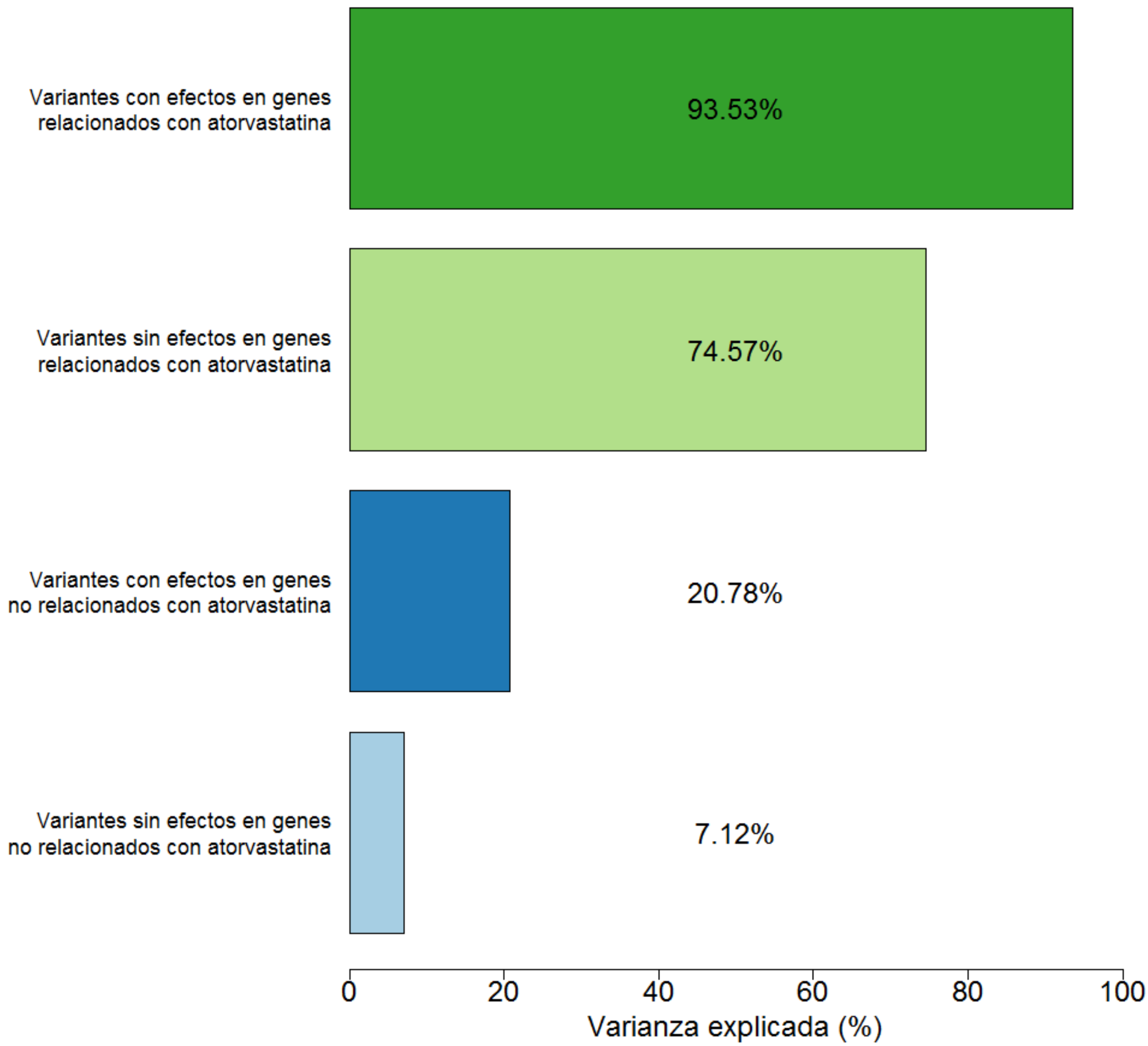
Para evaluar la validez del modelo de regresión refinado comparamos la varianza explicada por este modelo final contra la que se obtiene con 10 millones de modelos construidos al seleccionar aleatoriamente un número igual de variantes (60 variantes) de entre aquellas que tienen efectos predichos o reportados y se encuentran en genes relacionados con atorvastatina. Este análisis reveló que el valor p para obtener un modelo con varianza explicada igual o mayor a aquella del modelo refinado final es menor que $1x10^{-7}$.

Finalmente, se repitieron todos los pasos para la construcción y refinación de un modelo de regresión con vectores de soporte para contrastar el desempeño de los modelos que consideran únicamente las 60 mejores variantes de grupos de variantes genéticas que difieren en sus posibles efectos funcionales y al grupo de genes donde se encuentran. La varianza explicada que se obtuvo al construir y refinar modelos utilizando variantes localizadas en los genes relacionados a atorvastatina fue de 93.53% y 74.57%, para las variantes con efectos funcionales y las que carecen de los mismos, respectivamente. Mientras tanto la varianza explicada de los modelos construidos a partir de variantes localizadas en genes no relacionados a atorvastatina fue de 20.78% y 7.12% para los grupos de variantes con efectos funcionales y sin ellos.

Este análisis muestra que la varianza explicada por los modelos construidos a partir de variantes con efectos funcionales predichos o reportados es mayor que aquellos modelos construidos a partir de variantes sin dichos efectos. Del mismo modo, la varianza explicada por los modelos que consideran variantes presentes en los 20 genes relacionados al metabolismo y respuesta de atorvastatina son mayores que aquellos construidos a partir de variantes genéticas localizadas en genes no relacionados a este fármaco.

Adicionalmente, se utilizó la distribución de componentes ancestrales para intentar corregir estos modelos de regresión, encontrándose que la adición de estas variables no modificaba de manera importante los modelos generados (presentando cambios en la varianza explicada menores al 0.75%). Estos resultados sugieren que el análisis directo de las variantes genéticas presentes en los individuos es más importante para la asociación con la farmacocinética de atorvastatina que el cálculo de los componentes ancestrales de los mismos.

Figura 6. Desempeño de los modelos de regresión construidos a partir de diferentes grupos de variantes.



Varianza explicada por modelos de regresión construidos a partir de diferentes grupos de variantes determinados de acuerdo a los efectos funcionales que presentan y los genes en los cuales se localizan.

8.3 Catálogo de variantes genéticas en 430 genes relacionados a medicamentos

Al secuenciar los 60 individuos caracterizados farmacocinéticamente se encontraron 24,363 variantes totales en el total de 430 genes incluidos en el panel de enriquecimiento, de las cuales el 86.88% (21,166) son variaciones en un único nucleótido, el 6.38% (1,555) son inserciones y el 6.74% (1,642) son deleciones. Del total de 24,363 variantes encontradas en los 60 individuos el 17.88% (4,357) tiene efectos funcionales predichos por al menos una de las herramientas de predicción utilizadas (3,094 variantes), o están reportadas como asociadas a algún fármaco en la base de datos de PharmGKB (1,263 variantes). Mientras tanto, cabe resaltar que el número de variantes novedosas encontradas en estos individuos es de 3,714 y representa el 15.24% de todas las variantes detectadas, mientras que el 84.76% (20,649 variantes) se encuentra ya reportado en dbSNP versión 144.

Como se mencionó anteriormente, para complementar la caracterización de la diversidad genética en los 430 genes importantes de farmacogenómica se agregaron los datos obtenidos de otros dos proyectos de secuenciación importantes llevados a cabo en el INMEGEN. Al agregar los datos provenientes de 968 exomas de individuos mexicanos y 94 genomas completos de individuos amerindios a los 60 individuos caracterizados para la farmacocinética de atorvastatina, el número total de variantes en estos genes resulta ser de 47,319. De este total, el 25.49% (12,060) tiene efectos funcionales reportados en PharmGKB o según al menos uno de los predictores. Es importante mencionar que de las variantes con efectos, el 88.55% (10,679 variantes) sólo presenta efectos predichos y el restante 11.45% (1,381 variantes) es el que presenta asociaciones a fármacos en PharmGKB. Por otra parte, cabe resaltar que el 19.63% (9,288) de todas las variantes es novedoso y no se encuentra reportado en dbSNP 144.

Tabla 1. Clasificación de las variantes genéticas en 430 genes relacionados a medicamentos.

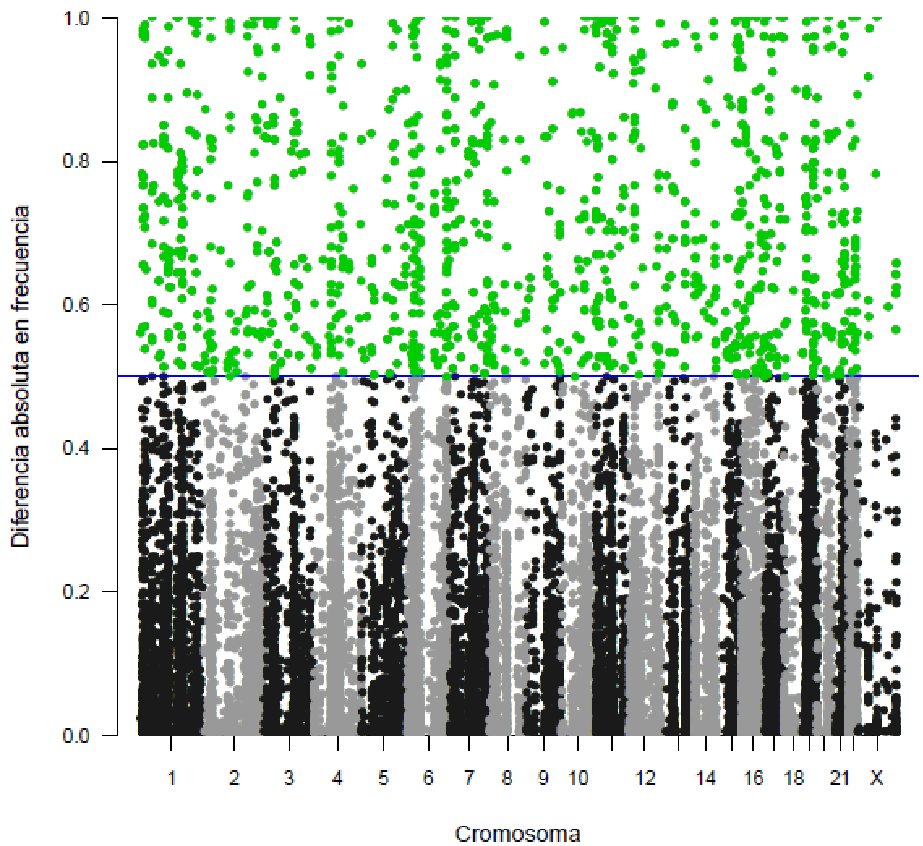
Categoría	**Tipo**	**Variantes totales**	**(%)**	**Variantes en 20 genes de atorvastatina**	**(%)**
Cambio de nucleótidos	Inserciones	1,642	(3.47)	59	(2.46)
	Deleciones	1,555	(3.29)	79	(3.29)
	SNVs	21,166	(44.73)	1,023	(42.61)
	MNVs	22,956	(48.51)	1,240	(51.65)
Novedad	Nuevas	9,288	(19.63)	377	(15.70)
	Reportadas	38,031	(80.37)	2,024	(84.30)
Efectos funcionales	Con efecto	12,060	(25.49)	758	(31.57)
	Sin efecto	35,259	(74.51)	1,643	(68.43)
Total		47,319	(100)	2,401	(100)

Clasificación de las variantes encontradas en todos los individuos (60 individuos caracterizados farmacocinéticamente para atorvastatina, 968 exomas de individuos mexicanos y 94 genomas de individuos amerindios).

Es importante resaltar que de todas las variantes el 5.07% (2,401 variantes) se encuentran localizadas en los 20 genes que intervienen en el metabolismo y respuesta ante atorvastatina, alcanzando aproximadamente el doble del total de variantes encontradas únicamente en los 60 individuos caracterizados farmacocinéticamente (1,161 variantes). De este total el 15.70% (377 variantes) son nuevas y el 31.57% (758 variantes) tienen efectos funcionales predichos (27.20%) o reportados (5.79%).

Por otro lado, al comparar la frecuencia con la que se encuentran las variantes en los individuos amerindios y aquella reportada en a nivel mundial en la base de datos de kaviar, encontramos 8,896 variantes que presentan una frecuencia significativamente diferente ($p < 1x10^{-8}$) según una prueba exacta de Fisher. Estas diferencias en frecuencia pueden ser muy grandes y de las variantes con diferencias significativa el 14.59% (1,298 variantes) presentan una diferencia en frecuencia de al menos 0.5 en valor absoluto.

Figura 7. Diferencia entre la frecuencia reportada para las variantes a nivel mundial y aquella que presentan en individuos amerindios



Diferencia absoluta entre la frecuencia a nivel mundial reportada en kaviar para cada variante y aquella con la que se encontró en los 94 individuos amerindios analizados. La línea marca el valor correspondiente a una diferencia en frecuencia de 0.5 en valor absoluto.

9. Discusión

9.1 Ancestría

Los componentes ancestrales amerindio, europeo y africano para los 60 individuos caracterizados farmacocinéticamente reclutados en Nuevo León, presentaron una mediana de 49.9%, 45.15% y 4.52% respectivamente. Esto concuerda con lo reportado anteriormente para otros estados del norte de México, específicamente Sonora y Tamaulipas, que en promedio tienen 43.90% amerindio, 52.02% europeo y 4.1% africano (72). Es de notar que las diferencias entre los resultados son pequeñas a pesar de que en este estudio se utilizaron únicamente 446 SNPs para analizar la ancestría, mientras que el análisis reportado para individuos de Sonora y Tamaulipas estuvo basado en 71,581 SNPs. Esta observación subraya la utilidad de incorporar marcadores genéticos relacionados al análisis de ancestría dentro de los paneles de secuenciación para obtener de forma simultánea el análisis de ancestría de los individuos y el descubrimiento de variantes genéticas asociadas a fenotipos de interés, aun cuando el número de marcadores analizados sea relativamente bajo.

El cálculo de las proporciones de ancestría europea, amerindia y africana a nivel individual presentó una importancia menor a la esperada, ya que no modificó en gran medida los modelos de regresión con vectores de soporte generados. Esto probablemente se debe a que dichos modelos fueron generados a partir de información sobre las variantes genéticas que se encuentran presentes realmente en los individuos, información que los componentes ancestrales pueden únicamente proporcionar de manera aproximada. Esto sugiere que para poder modelar la farmacocinética de atorvastatina es mejor contar con las variantes genéticas presentes en el individuo, sin embargo, no descarta la utilidad del cálculo de ancestría en ausencia de esta información o su importancia en el modelado de otros parámetros de efecto farmacológico de este u otros fármacos.

9.2 Farmacocinética de atorvastatina

El modelo refinado de regresión que utiliza únicamente las 60 variantes más importantes muestra una buena correlación con los valores observados de AUC (R^2=0.884). Sin embargo, los 4 individuos que presentan mayor AUC son los representados menos confiablemente por el modelo. Un análisis adicional, exhaustivo y cuidadoso de estos individuos podría revelar otros factores desconocidos hasta el momento –genéticos o de otra clase– que modifican la farmacocinética de atorvastatina.

Es importante notar que este modelo final de regresión incluye variantes previamente descritas como importantes en la farmacocinética de atorvastatina; por ejemplo, el alelo C de la variante rs4149056 en *SLCO1B1*, la cual ha sido asociada a un incremento de 144% en AUC de atorvastatina en individuos homocigotos (35). Sin embargo, ninguna variante logró explicar la mayor parte de la farmacocinética de atorvastatina por sí sola y para lograrlo fue necesario analizar la presencia de variantes genéticas a lo largo de varios genes que intervienen en todas las fases del metabolismo de atorvastatina.

La definición de un grupo de genes involucrados en el metabolismo y la respuesta ante atorvastatina que se analizaron tiene cierto grado de arbitrariedad y está restringida a genes incluidos en el experimento de secuenciación original. Sin embargo, resultados preliminares de este trabajo indican que el incluir un mayor número de genes con asociaciones reportadas con el perfil farmacocinético o farmacodinámico de un medicamento específico mejora el modelo de regresión en términos de la proporción de varianza explicada por éste.

Por su parte, la validación interna del modelo muestra que es posible construir un mejor modelo a partir de las variantes con efectos funcionales predichos o reportado que se encuentran en los genes relacionados con el fármaco estudiado (varianza explicada de 93.35%), que a partir de otros grupos de variantes. De manera importante este análisis mostró también que el modelo que se obtiene a partir de variantes sin efectos funcionales que se localizan en los 20 genes relacionados a atorvastatina explica una proporción mayoritaria de la varianza (74.57%), probablemente debido a que presentan de desequilibrio de unión con otras variantes funcionales cercanas. Por otra parte, el modelo construido a partir de variantes con efectos funcionales que están localizadas en genes no relacionados con atorvastatina explica una proporción de la varianza menor pero no despreciable (20.78%), lo cual resalta el hecho de que pueden existir genes que pueden afectar la farmacocinética de atorvastatina a través de vías desconocidas hasta el momento y que por lo tanto no están consideradas dentro del grupo de 20 genes relacionados a atorvastatina. Además, se mostró que el modelo construido a partir de las variantes sin efectos funcionales que se encuentran en genes no relacionados con atorvastatina explica una mucho menor proporción de la varianza (7.12%), lo cual concuerda con lo que habría de esperarse.

De acuerdo a nuestro análisis, para lograr predecir el perfil farmacocinético de atorvastatina –representado por el parámetro AUC– se necesita un análisis integral de diferentes genes que se sabe intervienen en la absorción, el metabolismo fase I y fase II, la excreción y la respuesta ante este fármaco (Anexo 3) y ninguna variante determina el perfil farmacocinético por sí misma.

Generalmente se acepta que los pasos limitantes de la farmacocinética de atorvastatina son los transportadores OATP1B1 –codificado en *SLCO1B1*–, glicoproteína p –codificada en *ABCB1*– o la proteína de resistencia de cáncer de mama –codificada en *ABCG2*. Esto esta soportado parcialmente por nuestros resultados, ya que el 21.66% de las variantes incluidas en el modelo refinado se encuentran en estos tres genes (13 variantes). Sin embargo, las dos variantes con mayor importancia en el modelo se localizan en el gen *CYP3A4*, el cual se encuentra involucrado en el metabolismo fase I de atorvastatina, demostrando una vez más la imperiosa necesidad de incluir las diferentes etapas en el metabolismo y respuesta de atorvastatina en el análisis.

El esquema presentado en este trabajo, donde un grupo de variantes genéticas priorizadas por su importancia relativa es utilizado para construir un modelo de regresión explicando la variabilidad en el AUC de atorvastatina, difiere de forma importante de otros estudios farmacogenómicos basados en secuenciación de nueva generación (73,74), ya que éstos tenían como objetivo la identificación de variantes específicas que presentaran un efecto sustancial en el fenotipo observado. Sin embargo, todos estos estudios subrayan la importancia de evaluar tanto variantes comunes como aquellas más raras, ya sean novedosas o previamente conocidas, para lograr explicar la variabilidad interpersonal de la respuesta ante fármacos.

El hecho de que en nuestro caso, ninguna variante fue capaz de explicar por sí sola la mayoría de la variabilidad en AUC de atorvastatina puede deberse a que algunas de las variantes que han sido previamente reportadas como determinantes de la variación en el metabolismo de atorvastatina fueron encontradas en frecuencias relativamente bajas y por lo tanto es necesario contar con otras variantes para explicar la variación interpersonal observada en la población estudiada. Por ejemplo, la variante rs2231142, que produce el cambio Gln141Lys en el transportador codificado en *ABCG2* y que ha sido asociada con mayores concentraciones plasmáticas de atorvastatina (29), se encontró en los 60 individuos con una frecuencia del 16.67%, frecuencia notablemente diferente de la reportada para las poblaciones europea (9%), asiática oriental (29%) y africana (1%) en la fase 3 del proyecto de 1,000 genomas (59). Del mismo modo, rs4149056, que produce el cambio Val174Ala en el transportador codificado en *SLCO1B1*, que se encuentra también asociada con una mayor concentración plasmática de atorvastatina (35) está presente con una frecuencia de sólo 10.83% en los 60 individuos, frecuencia menor a la reportada para la población europea (16%), menor pero similar a la reportada para la población asiática oriental (12%) y notablemente diferente de la reportada para la población africana (1%) del proyecto de 1,000 genomas (59).

También se observa el caso contrario, en el que las variantes son tan frecuentes que es necesario contar con variantes adicionales para explicar la variabilidad observada en la farmacocinética. De esta forma otras variantes que también han sido consistentemente consideradas como importantes en el área de la farmacogenómica fueron mucho más frecuentes, tanto que se encuentran presentes en casi todos los individuos. Por ejemplo, la variante rs2740574, un cambio -392 A>G en la región 5' de CYP3A4 que incrementa la transcripción de esta enzima de la familia citocromo P450 y por lo tanto su participación en el metabolismo fase I de la atorvastatina se encontró en la población estudiada con una frecuencia de 92.5%, frecuencia menor pero relativamente similar a la reportada para la poblaciones europea (97%) y asiática oriental (99.6%) y notablemente diferente de la reportada para la población africana (23%) en el proyecto de 1000 genomas.. En este aspecto es importante mencionar que las 60 variantes incluidas en el modelo final de regresión presentan frecuencias alélicas con una mediana de 35.42% (RI: 13.23 – 52.46%), aunque cinco variantes presentan frecuencias alélicas menores a 5% por lo que pueden considerarse como raras para esta población.

A pesar de contar con un tamaño de muestra relativamente pequeño (n=60), nuestros resultados sugieren que la mayor parte de la variación en la farmacocinética de atorvastatina puede ser explicada utilizando un conjunto de marcadores genéticos soportados por la predicción del efecto funcional de cada variante. Sin embargo, una evaluación del modelo de regresión en una muestra independiente de la población es el único método para validar adecuadamente el desempeño del modelo en nuevos individuos.

En este momento, es imposible evaluar el grado en el cual este modelo podrá aplicarse a poblaciones con un contexto genético diferente del considerado durante la construcción del mismo –en este caso voluntarios sanos del noreste de México–, y para lograrlo se necesitaría una evaluación rigurosa en poblaciones diferentes, incluyendo aquellas pertenecientes a diferentes regiones geográficas del país.

Como se mencionó con anterioridad en este trabajo se generaron los modelos de regresión utilizando vectores de soporte debido principalmente a que son capaces de manejar el número de muestras relativamente pequeño con el cual se contaba. Sin embargo, otros algoritmos de aprendizaje automatizado podrían ser más adecuados en otros casos, al proveer ventajas que los vectores de soporte no tienen, por ejemplo, se podrían utilizar redes neuronales artificiales cuando se cuente con un gran volumen de muestras de entrenamiento gracias a su capacidad de implementarse eficientemente en arquitecturas de cómputo en paralelo.

Por otro lado, aun cuando se logró construir un modelo de regresión explicando la mayor parte de la variabilidad en AUC de atorvastatina a partir de un grupo de variantes genéticas el reto de descubrir el mecanismo por el cual cada una de ellas ejerce su efecto sobre la farmacocinética de este fármaco todavía persiste. Idealmente se requeriría proyectar el efecto biológico de cada una de las variantes sobre los genes participantes en el metabolismo y respuesta de atorvastatina y evaluar cómo a través de las interacciones entre ellos se obtiene el efecto total observado sobre los niveles plasmáticos de atorvastatina. Sin embargo, nuestra exploración preliminar de este aspecto indica que es demasiado complejo para llevarlo a cabo dentro del marco de este trabajo.

En nuestro caso y como sucede en todos los estudios donde se relaciona la presencia de variantes genéticas con un fenotipo específico, algunas de las variantes incluidas en el modelo de regresión pueden no ser las responsables directas del efecto sobre la farmacocinética de atorvastatina. Como consecuencia, para lograr determinar si estas variantes tienen un efecto directo o si son marcadores relacionados estrechamente a otras variantes funcionales subyacentes, se necesitaría emprender estudios funcionales, aún si el uso de las herramientas informáticas de predicción de efecto deberían disminuir la probabilidad de observar este efecto.

Una barrera importante para la implementación de un análisis integral de muchos genes relevantes –no sólo para el metabolismo y respuesta de atorvastatina, sino también para un amplio espectro de otros fármacos– en la práctica clínica actual, es el alto costo asociado con los experimentos de secuenciación de nueva generación utilizados. Sin embargo, la disminución del costo de las tecnologías de genotipificación, el aumento de la cantidad de marcadores farmacogenómicos validados y el ahora creciente número de individuos genotipificados de forma preventiva (75) son factores que contribuyen a aumentar la utilidad de los enfoques similares al utilizado en este trabajo y favorecen su futura implementación en la práctica clínica.

9.3 Catálogo de variantes genéticas en 430 genes relacionados a medicamentos

Como se mencionó anteriormente en este trabajo se compiló un catálogo de 47,319 variantes genéticas presentes en más de 400 genes relacionados al metabolismo y respuesta de diversos fármacos, encontradas en 1,122 individuos pertenecientes a la población mexicana.

Como se esperaba muchas de estas variantes presentan notables diferencias en frecuencia principalmente entre individuos amerindios y aquellas reportadas a nivel mundial, hecho que resalta la importancia de recopilar la variabilidad genética en individuos descendientes de diferentes grupos y regiones geográficas de México.

Una proporción importante (25.49%) de todas las variantes catalogadas presenta efectos funcionales por lo cual podrían modificar la respuesta farmacológica de diversos fármacos. Éstas son 12,060 variantes de las cuales 1,847 son nuevas y las restantes 10,213 habían sido descritas previamente. Únicamente 1,381 de éstas variantes tienen efectos reportados en PharmGKB y la gran mayoría de las variantes sólo tienen efectos funcionales predichos, lo cual resalta la importancia de utilizar diversos algoritmos y herramientas de predicción para evaluar el efecto de las variantes.

El número de variantes que todavía faltan por descubrir en las regiones de farmacogenómica analizadas se espera que sea relativamente bajo, ya que al considerar que un aumento en el tamaño de muestra mayor a 17 veces (de 60 a un total de 1,122 individuos incluidos en este trabajo) tuvo como consecuencia un aumento del número de variantes encontradas aproximadamente al doble podemos suponer que el número de variantes se encuentra cerca de la saturación. Sin embargo, el análisis de nuevos individuos principalmente amerindios pertenecientes a diferentes regiones geográficas de México podría ayudar a complementar la variación genética descrita hasta el momento.

Tal como se demostró para el caso de la farmacocinética de atorvastatina, la importancia de este grupo de 12,060 variantes con efectos funcionales predichos o reportados radica en su potencial para describir (de mejor manera) el metabolismo y respuesta de diversos fármacos en la población mexicana. Por lo tanto estas variantes podrían y deberían considerarse para realizar los diseños preliminares de plataformas de genotipificación farmacogenómica dirigidas a aumentar la efectividad terapéutica y reducir la aparición de reacciones adversas en la población mexicana.

10. Conclusiones

La diferencia en metabolismo y eficacia de los fármacos dependen de una gran variedad de factores, tanto individuales como ambientales. Se sabe que variantes genéticas, las cuales pueden ser específicas de una población determinada, pueden ejercer un efecto sobre los diferentes pasos de metabolismo, transporte y excreción de un mismo fármaco. Por esta razón resulta necesario conocer la diversidad genética presente en los genes que intervienen en diferentes etapas del metabolismo de diversos fármacos para poder asociar de mejor manera los factores genéticos al metabolismo y la respuesta ante diversos fármacos.

Se logró utilizar las variantes con efectos funcionales predichos o reportados que se encuentran presentes en 20 genes (*ABCA1*, *ABCB1*, *ABCC2*, *ABCG2*, *ACE*, *CYP2D6*, *CYP3A4*, *CYP3A5*, *HMGCR*, *LEPR*, *NOS3*, *POR*, *SLCO1B1*, *SLCO1B3*, *SLCO2B1*, *TNF*, *UGT1A1*, *UGT1A3*, *UGT1A4* y *UGT2B7*) para representar fielmente el fenómeno de la variabilidad farmacocinética de atorvastatina en 60 individuos sanos del noreste de México. Este resultado clave demuestra la importancia de la caracterización de los genes importantes en farmacogenómica para permitir la asociación a fenotipos de interés para la salud pública de nuestro país, como son la farmacocinética y la respuesta ante fármacos.

Estos resultados proporcionan evidencia que soporta el uso de una combinación de herramientas computacionales para realizar una predicción del efecto funcional de las variantes todavía no caracterizadas experimentalmente. Esta predicción de efecto funcional mediante diversas herramientas es fundamental para evaluar en un primer nivel la posible importancia de las variantes novedosas encontradas mediante esquemas de secuenciación de nueva generación como el que fue utilizado en este proyecto.

Además, los resultados soportan el uso de enfoques integrales, donde se consideren simultáneamente las variantes presentes en diversos genes que intervienen en las diferentes etapas del metabolismo de un mismo fármaco para predecir de mejor manera su efecto sobre el fenotipo de interés, en este caso la farmacocinética de atorvastatina.

Los resultados de este trabajo también permiten caracterizar la diversidad genética de un grupo de más de 400 genes importantes en farmacogenómica en una muestra de la población mexicana. Esta muestra, al encontrarse enriquecida con la información proveniente de otros dos diferentes proyectos de secuenciación que en conjunto comprenden más de mil individuos mexicanos (de los cuales aproximadamente 100 son individuos amerindios) es la caracterización farmacogenómica enfocada a la población de México más completa hasta el momento.

El utilizar tecnologías de secuenciación masiva permitió no solo evaluar la presencia de las variantes previamente identificadas como asociadas al metabolismo y respuesta ante diversos fármacos, sino al mismo tiempo también el descubrimiento de un gran número de variantes novedosas que se encuentran en más de 400 genes importantes para el área de la farmacogenómica. Como consecuencia de este estudio se ha obtenido una lista de variantes con posible efecto funcional que pueden afectar el metabolismo y respuesta a fármacos que puede utilizarse de forma preliminar para el diseño de plataformas de genotipificación masiva (por ejemplo microarreglos de ADN) que permitirán la realización de estudios poblacionales sobre la frecuencia –y posible asociación con respuesta a fármacos– de estas variantes a nivel nacional.

11. PERSPECTIVAS

Al utilizar la farmacogenómica para predecir la efectividad o toxicidad de algunos fármacos en la práctica cínica actual es posible observar dos casos. Por un lado, unas cuantas variantes de gran penetrancia permiten predecir la efectividad o toxicidad de un fármaco con suficiente confianza para realizar una indicación terapéutica, mientras que por el otro, las variantes genéticas conocidas presentan efectos tan pequeños –o fenotipos tan complejos– que no es posible realizar recomendación alguna. En este segundo grupo de fármacos –en el cual se incluye a la atorvastatina– es donde serán más útiles los esquemas novedosos donde se analicen la totalidad de las variantes genéticas relevantes para el metabolismo y respuesta a fármacos, permitiendo proveer indicaciones terapéuticas donde actualmente no es posible hacerlo a través del análisis de una o dos variantes.

En este trabajo se generó un modelo de regresión que utiliza variantes genéticas presentes en genes que intervienen en todas las etapas del metabolismo y respuesta de atorvastatina para explicar la mayor parte de la variabilidad interpersonal en cuanto a la farmacocinética de este fármaco. Sin embargo, todavía es necesario evaluar el modelo de regresión generado en muestras independientes provenientes de otras regiones geográficas del país para evaluar el potencial de generalización del modelo. Esto requeriría realizar la evaluación de los parámetros farmacocinéticos para este fármaco y la genotipificación de al menos las 60 variantes incluidas en el modelo de regresión en individuos provenientes de diferentes regiones de México o de otros países.

Por otro lado, también sería conveniente evaluar los méritos de la estrategia seguida en este trabajo para modelar la respuesta en reducción de lípidos ante un tratamiento con atorvastatina y no sólo acerca de la farmacocinética de este medicamento. Además, esto podría evaluarse utilizando poblaciones de individuos con hipercolesterolemia en vez de individuos saludables, tomando en cuenta como covariables del modelo otros factores no considerados en este trabajo como la presencia de enfermedades concomitantes y el consumo de otros medicamentos. Esto ayudaría a establecer la utilidad real de modelos similares en la práctica clínica actual y propiciaría su extensión a diferentes patologías y medicamentos.

Además, se obtuvo una lista de 12,060 variantes genéticas que presentan efectos funcionales predichos o reportados sobre el metabolismo de fármacos las cuales pueden factiblemente ser analizadas simultáneamente mediante el uso de plataformas de genotipificación comerciales. De hecho, esta lista ha comenzado ya a refinarse para eliminar las variantes menos útiles desde el punto de vista de salud pública como por ejemplo las variantes privadas –presentes en únicamente un individuo– para incorporarse al diseño de microarreglos de genotipificación dirigidos específicamente a la población mexicana.

Este tipo de plataformas generales vuelve factible hoy día la genotipificación de este número de variantes a un costo razonable, permitiendo por ejemplo, su utilización en estudios sobre la efectividad y toxicidad de fármacos de especial interés para México. De este modo, todos los individuos a los que se genotipifique utilizando esta plataforma general contarán adicionalmente con la información relevante para todos los demás fármacos que puedan prescribirles a futuro sin necesidad de repetir el estudio.

Con esto en mente es posible generar esquemas aún más útiles para la práctica clínica al combinar la genotipificación de las variantes de farmacogenómica con otras variantes importantes, por ejemplo las que están relacionadas con la aparición de enfermedades que presentan patrones de herencia mendeliana y evaluar su presencia simultáneamente desde el nacimiento (o incluso antes).

12. Referencias

1. Spear BB, Heath-Chiozzi M, & Huff J (2001) Clinical application of pharmacogenetics. *Trends Mol Med* 7(5):201-204.
2. Impicciatore P*, et al.* (2001) Incidence of adverse drug reactions in paediatric in/out-patients: a systematic review and meta-analysis of prospective studies. *British Journal of Clinical Pharmacology* 52(1):77-83.
3. Eichelbaum M, Ingelman-Sundberg M, & Evans WE (2006) Pharmacogenomics and individualized drug therapy. *Annu Rev Med* 57:119-137.
4. Clinical Pharmacogenetics Implementation Consortium (2016) CPIC Guidelines. Genes-Drugs.
5. Bustamante CD, Burchard EG, & De la Vega FM (2011) Genomics for the world. *Nature* 475(7355):163-165.
6. Vargens DD, Damasceno A, Petzl-Erler ML, & Suarez-Kurtz G (2011) Combined CYP2C9, VKORC1 and CYP4F2 frequencies among Amerindians, Mozambicans and Brazilians. *Pharmacogenomics* 12(6):769-772.
7. Nelson MR*, et al.* (2012) An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* 337(6090):100-104.
8. Cuautle-Rodriguez P, Llerena A, & Molina-Guarneros J (2014) Present status and perspective of pharmacogenetics in Mexico. *Drug Metabol Drug Interact* 29(1):37-45.
9. Contreras AV*, et al.* (2011) Resequencing, haplotype construction and identification of novel variants of CYP2D6 in Mexican Mestizos. *Pharmacogenomics* 12(5):745-756.
10. Kocabas NA*, et al.* (2008) Gemcitabine pharmacogenomics: Deoxycytidine kinase and cytidylate kinase gene resequencing and functional genomics. *Drug Metabolism and Disposition* 36(9):1951-1959.
11. Martin YN*, et al.* (2006) Human methylenetetrahydrofolate reductase pharmacogenomics: gene resequencing and functional genomics. *Pharmacogenetics and Genomics* 16(4):265-277.
12. Licinio J, Dong C, & Wong ML (2009) Novel sequence variations in the brain-derived neurotrophic factor gene and association with major depression and antidepressant treatment response. *Arch Gen Psychiatry* 66(5):488-497.
13. Dong C, Wong ML, & Licinio J (2009) Sequence variations of ABCB1, SLC6A2, SLC6A3, SLC6A4, CREB1, CRHR1 and NTRK2: association with major depression and antidepressant response in Mexican-Americans. *Molecular Psychiatry* 14(12):1105-1118.
14. Chen Y*, et al.* (2009) Genetic variants in multidrug and toxic compound extrusion-1, hMATE1, alter transport function. *Pharmacogenomics Journal* 9(2):127-136.
15. National Human Genome Research Institute (2013) DNA sequencing costs. Data from the NHGRI large-scale genome sequencing program.
16. Ng SB*, et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461(7261):272-U153.
17. Cooper GM & Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* 12(9):628-640.
18. Fernandez G, Spatz ES, Jablecki C, & Phillips PS (2011) Statin myopathy: a common dilemma not reflected in clinical trials. *Cleveland Clinic journal of medicine* 78(6):393-403.
19. Lennernas H (2003) Clinical pharmacokinetics of atorvastatin. *Clin Pharmacokinet* 42(13):1141-1160.
20. Chen CP*, et al.* (2005) Differential interaction of 3-hydroxy-3-methylglutaryl-COA reductase inhibitors with ABCB1, ABCC2, and OATP1B1. *Drug Metabolism and Disposition* 33(4):537-546.
21. Karlgren M*, et al.* (2012) Classification of Inhibitors of Hepatic Organic Anion Transporting Polypeptides (OATPs): Influence of Protein Expression on Drug - Drug Interactions. *Journal of Medicinal Chemistry* 55(10):4740-4763.
22. Park JE*, et al.* (2008) Contribution of cytochrome P450 3A4 and 3A5 to the metabolism of atorvastatin. *Xenobiotica* 38(9):1240-1251.
23. Goosen TC*, et al.* (2007) Atorvastatin glucuronidation is minimally and nonselectively inhibited by the fibrates gemfibrozil, fenofibrate, and fenofibric acid. *Drug Metabolism and Disposition* 35(8):1315-1324.

24. Prueksaritanont T*, et al.* (2002) Glucuronidation of statins in animals and humans: A novel mechanism of statin lactonization. *Drug Metabolism and Disposition* 30(5):505-512.
25. Voora D*, et al.* (2008) Pharmacogenetic Predictors of Statin-Mediated Low-Density Lipoprotein Cholesterol Reduction and Dose Response. *Circulation-Cardiovascular Genetics* 1(2):100-106.
26. Potoczek DP, Undas A, Iwaniec T, & Szczeklik A (2005) The angiotensin-converting enzyme gene insertion/deletion polymorphism and effects of quinapril and atorvastatin on haemostatic parameters in patients with coronary artery disease. *Thrombosis and Haemostasis* 94(1):224-225.
27. Donnelly LA*, et al.* (2008) A paucimorphic variant in the HMG-CoA reductase gene is associated with lipid-lowering response to statin treatment in diabetes: a GoDARTS study. *Pharmacogenet Genomics* 18(12):1021-1026.
28. Chien KL*, et al.* (2010) Common sequence variants in pharmacodynamic and pharmacokinetic pathway-related genes conferring LDL cholesterol response to statins. *Pharmacogenomics* 11(3):309-317.
29. Chung JY*, et al.* (2012) Effect of HMGCR variant alleles on low-density lipoprotein cholesterol-lowering response to atorvastatin in healthy Korean subjects. *J Clin Pharmacol* 52(3):339-346.
30. Cuevas A*, et al.* (2016) HMGCR rs17671591 SNP Determines Lower Plasma LDL-C after Atorvastatin Therapy in Chilean Individuals. *Basic & Clinical Pharmacology & Toxicology* 118(4):292-297.
31. Drogari E*, et al.* (2014) POR*28 SNP is associated with lipid response to atorvastatin in children and adolescents with familial hypercholesterolemia. *Pharmacogenomics* 15(16):1963-1972.
32. Cho SK, Oh ES, Park K, Park MS, & Chung JY (2012) The UGT1A3*2 polymorphism affects atorvastatin lactonization and lipid-lowering effect in healthy volunteers. *Pharmacogenetics and Genomics* 22(8):598-605.
33. Willrich MA*, et al.* (2008) CYP3A53A allele is associated with reduced lowering-lipid response to atorvastatin in individuals with hypercholesterolemia. *Clin Chim Acta* 398(1-2):15-20.
34. Becker ML*, et al.* (2010) Influence of genetic variation in CYP3A4 and ABCB1 on dose decrease or switching during simvastatin and atorvastatin therapy. *Pharmacoepidemiology and Drug Safety* 19(1):75-81.
35. Ulvestad M*, et al.* (2013) Impact of OATP1B1, MDR1, and CYP3A4 Expression in Liver and Intestine on Interpatient Pharmacokinetic Variability of Atorvastatin in Obese Subjects. *Clinical Pharmacology & Therapeutics* 93(3):275-282.
36. Keskitalo JE*, et al.* (2009) ABCG2 Polymorphism Markedly Affects the Pharmacokinetics of Atorvastatin and Rosuvastatin. *Clinical Pharmacology & Therapeutics* 86(2):197-203.
37. Becker ML*, et al.* (2013) Genetic variation in the ABCC2 gene is associated with dose decreases or switches to other cholesterol-lowering drugs during simvastatin and atorvastatin therapy. *Pharmacogenomics Journal* 13(3):251-256.
38. Frudakis TN*, et al.* (2007) CYP2D6*4 polymorphism is associated with statin-induced muscle effects. *Pharmacogenetics and Genomics* 17(9):695-707.
39. Perez-Castrillon JL*, et al.* (2009) Atorvastatin and BMD in Coronary Syndrome. Role of Lys656Asn Polymorphism of Leptin Receptor Gene. *Endocrine Journal* 56(2):221-225.
40. Perez-Castrillon JL*, et al.* (2008) Effect of the TNF alpha-308 G/A Polymorphism on the Changes Produced by Atorvastatin in Bone Mineral Density in Patients with Acute Coronary Syndrome. *Annals of Nutrition and Metabolism* 53(2):117-121.
41. Link E*, et al.* (2008) SLCO1B1 variants and statin-induced myopathy - A genomewide study. *New England Journal of Medicine* 359(8):789-799.
42. Amirimani B*, et al.* (2003) Increased transcriptional activity of the CYP3A4*1B promoter variant. *Environ Mol Mutagen* 42(4):299-305.
43. Kuehl P*, et al.* (2001) Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nature Genetics* 27(4):383-391.
44. Kivisto KT*, et al.* (2004) Lipid-lowering response to statins is affected by CYP3A5 polymorphism. *Pharmacogenetics* 14(8):523-525.

45. Morisaki K*, et al.* (2005) Single nucleotide polymorphisms modify the transporter activity of ABCG2. *Cancer Chemotherapy and Pharmacology* 56(2):161-172.
46. Williams AL*, et al.* (2014) Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* 506(7486):97-+.
47. Leon-Cachon RBR*, et al.* (2015) A pharmacogenetic pilot study reveals MTHFR, DRD3, and MDR1 polymorphisms as biomarker candidates for slow atorvastatin metabolizers. *BMC cancer* 16:74-74.
48. Affymetrix (2012) Get annotations for up to 10000 Affymetrix probe set accesion numbers, gene names or sequences ids.
49. Whirl-Carrillo M*, et al.* (2012) Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 92(4):414-417.
50. Galanter JM*, et al.* (2012) Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet* 8(3):e1002554.
51. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
52. McKenna A*, et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20(9):1297-1303.
53. Van der Auwera GA*, et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 43:11.10.11-33.
54. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19(9):1655-1664.
55. Choi Y, Sims GE, Murphy S, Miller JR, & Chan AP (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. *Plos One* 7(10).
56. McLaren W*, et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26(16):2069-2070.
57. Shihab HA*, et al.* (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31(10):1536-1543.
58. Glusman G, Caballero J, Mauldin DE, Hood L, & Roach JC (2011) Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 27(22):3216-3217.
59. Vapnik V, Golowich SE, & Smola A (1997) Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*:281-287.
60. Cristianini N & Shawe-Taylor J (2000) *An introduction to support vector machines and other kernel-based learning methods* (Cambridge university press).
61. Usdun B, Melssen WJ, & Buydens LMC (2007) Visualisation and interpretation of Support Vector Regression models. *Analytica Chimica Acta* 595(1-2):299-309.
62. Pedregosa F*, et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825-2830.
63. Cruz-Correa OF, León-Cachón RBR, Barrera-Saldaña HA, & Soberón X (2017) Prediction of atorvastatin plasmatic concentrations in healthy volunteers using integrated pharmacogenetics sequencing. *Pharmacogenomics* 18(2):121-131.
64. Moreno-Estrada A*, et al.* (2014) HUMAN GENETICS The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* 344(6189):1280-1285.
65. Pasanen MK, Fredrikson H, Neuvonen PJ, & Niemi M (2007) Different effects of SLCO1B1 polymorphism on the pharmacokinetics of atorvastatin and rosuvastatin. *Clinical Pharmacology & Therapeutics* 82(6):726-733.
66. Mizzi C*, et al.* (2014) Personalized pharmacogenomics profiling using whole-genome sequencing. *Pharmacogenomics* 15(9):1223-1234.

67. Tavira B*, et al.* (2015) ABCB1 (MDR-1) pharmacogenetics of tacrolimus in renal transplanted patients: a Next Generation Sequencing approach. *Clinical Chemistry and Laboratory Medicine* 53(10):1515-1519.
68. Hoffman JM*, et al.* (2014) PG4KDS: A model for the clinical implementation of pre-emptive pharmacogenetics. *American Journal of Medical Genetics Part C-Seminars in Medical Genetics* 166(1):45-55.

Anexo 1. Artículo derivado del trabajo de investigación.

Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Pharmacogenomics

Prediction of atorvastatin plasmatic concentrations in healthy volunteers using integrated pharmacogenetics sequencing

Aim: To use variants found by next-generation sequencing to predict atorvastatin plasmatic concentration profiles (AUC) in healthy volunteers. **Subjects & methods:** A total of 60 healthy Mexican volunteers were enrolled in this study. We used variants with a predicted functional effect across 20 genes involved in atorvastatin metabolism to construct a regression model using a support vector approach with a radial basis function kernel to predict AUC refining it afterwards in order to explain a greater extent of the variance. **Results:** The final support vector regression model using 60 variants (including six novel variants) explained 94.52% of the variance in atorvastatin AUC. **Conclusion:** An integrated analysis of several genes known to intervene in the different steps of metabolism is required to predict atorvastatin's AUC.

First draft submitted: 21 April 2016; Accepted for publication: 29 September 2016; Published online: 15 December 2016

Keywords: atorvastatin • Next-Gen sequencing • pharmacokinetics

Omar Fernando Cruz-Correa[1], Rafael Baltazar Reyes León-Cachón[2,3], Hugo Alberto Barrera-Saldaña[2,4] & Xavier Soberón*[1,5]

[1]Instituto Nacional de Medicina Genómica, Periférico Sur No. 4809, Col. Arenal Tepepan, Delegación Tlalpan, México, D.F. C.P. 14610, Mexico

[2]Departamento de Bioquímica y Medicina Molecular, Facultad de Medicina, Universidad Autónoma de Nuevo León, Ave. Madero, Col. Mitras Centro, Monterrey, Nuevo León, C.P. 64640, Mexico

[3]División Ciencias de la Salud, Departamento de Ciencias Básicas, Centro de Diagnóstico Molecular y Medicina Personalizada, Universidad de Monterrey, Ave. Ignacio Morones Prieto Pte. 4500, Col. Jesús M. Garza, San Pedro Garza García, Nuevo León, C.P. 66238, Mexico

[4]Vitagénesis, SA de CV., Col. Colinas de San Jerónimo. Monterrey, Nuevo León, C.P. 64630, Mexico

[5]Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, Cuernavaca, Morelos, C.P. 62210, Mexico

*Author for correspondence: xsoberon@inmegen.gob.mx

In recent years, pharmacogenetics has received much attention because of its potential to attain higher drug efficacy and lower adverse reactions through the personalization of drug therapy regimes [1] and to date (September 2016) there are 41 freely available peer-reviewed guidelines indicating how to use the genetic information on certain genes and drugs to achieve this optimization [2].

However, standardized pharmacogenetic dosing algorithms may have poorer performance when used in different populations, in these cases the known variants fail to explain to a full extent the interpersonal drug response or pharmacokinetic variability observed. This may be mainly to the presence of other variants specific to the studied population located in the same genes – or in other genes related to different steps in drug metabolism, transport, excretion and toxicity – which might also have an effect on the pharmacokinetics of the same drug [3].

Atorvastatin is a cholesterol-LDL lowering drug used in hypercholesterolemia patients to prevent progression to atherosclerosis and the ensuing cardiovascular risk. It belongs to a family of HMG-CoA reductase inhibitors – encoded in the *HMGCR* gene – commonly known as statins. Although this family of drugs is considered relatively safe, up to 20% of patients treated with statins may present a myopathy related to treatment where symptoms range from mild fatigue and muscular pain to potentially fatal rhabdomyolysis [4].

Atorvastatin is administered as a calcium salt, its active open form, with a dosage ranging from 10 to 80 mg once a day. In physiological conditions, the open form is in equilibrium with its lactone (inactive) form, and both the open and lactone forms present similar values of area under the plasma concentration curve (AUC) [5].

Once administered, atorvastatin transport depends upon several different gene products

Future Medicine part of fsg

including the organic anion transporting polypeptides (OATPs) encoded in the *SLCO1B1*, *SLCO1B3* and *SLCO2B1* genes, each presenting variable affinities for the open and lactone form [6,7].

Atorvastatin is metabolized to two pharmacologically active metabolites (2-hydroxi-atorvastatin and 4-hydroxi atorvastatin) and their corresponding inactive lactones, mainly in the liver by the action of the cytochrome encoded in *CYP3A4*, and, to a lesser extent, by the one encoded in *CYP3A5* [5,8].

These compounds are subject to glucuronidation by the action of several UDP glucuronosyltransferases. Although the most important for atorvastatin metabolism are the ones encoded in *UGT1A1*, *UGT1A3* and *UGT1A4*, it has been reported that UGT2B7 also has a small glucuronidation activity for this drug [9,10]. Atorvastatin elimination is carried out mainly through bile by the action of the transporters P-glycoprotein and BCRP, encoded respectively, in *ABCB1* and *ABCG2*.

Several genetic polymorphisms present in genes related to atorvastatin transport, distribution, metabolism, response and excretion modify an ensemble of different parameters, such as the rates of the lipid lowering effects of atorvastatin (rs12003906 in *ABCA1* [11], rs1799752 deletion polymorphism in *ACE* [12], rs17238540 [13], rs12916 [14], rs3846662 [15], rs17671591 [16] in *HMGCR*, rs1057868 in *POR* [17], rs8175347 in *UGT1A3* [18], rs776746 in *CYP3A5* [19]), its pharmacokinetic profile (rs2740574 in *CYP3A4* [20], rs4149056 [Val174Ala] in *SLCO1B1* [21]), rs2231142 in *ABCG2* [22] and the apparition of adverse reactions (rs717620 in *ABCC2* [23], rs3892097 in *CYP2D6* [24]). Also, there are reports of polymorphisms associated with effects of atorvastatin other than its lipid-lowering capabilities, which could relate to atorvastatin pharmacokinetics through not yet completely understood pathways (e.g., variants rs1805094 in *LEPR* [25] and rs1800629 in *TNF* [26] are associated with bone mineral density increments during atorvastatin treatment).

As atorvastatin pharmacokinetics parameters may be modified by the ensemble of genetic variants present in any of the genes involved in the different steps of atorvastatin metabolism or response, we used variants in the 20 previously mentioned genes (*ABCA1*, *ABCB1*, *ABCC2*, *ABCG2*, *ACE*, *CYP2D6*, *CYP3A4*, *CYP3A5*, *HMGCR*, *LEPR*, *NOS3*, *POR*, *SLCO1B1*, *SLCO1B3*, *SLCO2B1*, *TNF*, *UGT1A1*, *UGT1A3*, *UGT1A4* and *UGT2B7*) known to be involved in atorvastatin metabolism and response to predict AUC values in 60 healthy Mexican volunteers using Support Vector Regression with a Radial Basis Function kernel.

Subjects & methods

Subject selection

A total of 60 apparently healthy male volunteers from northeastern Mexico, 18–45 years of age, seronegative for HIV, HBV and HCV, with a BMI between 20 and 26 kg/m^2 and a normal complete blood count, blood chemistry and urinalysis, participating in a pharmacokinetics study of atorvastatin (Australian New Zealand Clinical Trials Registry ACTRN12614000851662) were enrolled in this study. Written informed consent was obtained from all volunteers and the protocol was approved by the Research and Ethics Committee of Mexico's National Institute of Genomic Medicine (INMEGEN, Mexico). DNA was extracted from whole-blood and subjected to enrichment and next-generation sequencing as described below.

Pharmacokinetic analysis

Pharmacokinetics profiling was performed as described elsewhere [27]. Briefly, after an overnight fast, individuals were given a single dose of 80 mg of atorvastatin calcium. Blood samples were drawn before atorvastatin administration and at 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 5.0, 6.0, 8.0, 12.0, 24.0, 36.0 and 48.0 h. Blood samples were centrifuged and plasma was separated and subjected to high performance liquid chromatography-tandem mass spectrometry (HPLC/MS/MS) to measure levels of atorvastatin in its open form. AUC from time 0 to time of last determination was estimated considering a noncompartmental method and using the linear-logarithmic trapezoidal method for calculation.

Next-generation library preparation & sequencing

A custom HaloPlex™ probe panel (Agilent Technologies, CA, USA) was designed to enrich more than 400 genes involved in drug metabolism, effect or toxicity. Target enrichment libraries for next-generation sequencing were prepared according to the HaloPlex protocol. Briefly, genomic DNA extracted from whole-blood was fragmented with restriction enzymes, and hybridized in solution with custom-designed biotinylated probes to circularize fragments encompassing target regions, and 'barcode' oligonucleotides for sample identification. Circularized fragments were purified using magnetic beads coated with streptavidin and PCR amplified to include barcode and sequencing adapters. Libraries concentration was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies). Sequencing was carried out in a Genome Analyzer IIx (Illumina Inc., CA, USA) with a read-length of 150 bp in paired-end format.

Sequence analysis

Paired-end reads were trimmed to remove sequencing adapters and regions of low quality with Trimmomatic software [28]. Target enrichment performance metrics were obtained from the program CalculateHsMetrics in the software package Picard (picard-tools-1.130 [29]), then variants were called according to GATK best practices with the sole difference that duplicate reads were not marked because specific fragment start and end positions are expected due to the enzymatic fragmentation step during the enrichment protocol. Briefly, reads passing quality filters were aligned to the human reference genome build hg19 using BWA-mem [30], were subject to realignment around indels, base quality recalibration, SNV and indel calling, joint genotyping and variant quality score recalibration with GATK's [31] version 3.3-0-g37228af, built-in tools and recommended parameters [32]. Variants were further analyzed for functional effects using PROVEAN [33] and ENSEMBLE's Variant Effect Predictor (VEP) [34].

Construction of support vector regression models

We retained all variants present in 20 genes related to atorvastatin response, metabolism and effects (*ABCA1*, *ABCB1*, *ABCC2*, *ABCG2*, *ACE*, *CYP2D6*, *CYP3A4*, *CYP3A5*, *HMGCR*, *LEPR*, *NOS3*, *POR*, *SLCO1B1*, *SLCO1B3*, *SLCO2B1*, *TNF*, *UGT1A1*, *UGT1A3*, *UGT1A4* and *UGT2B7*) that were classified either as 'deleterious' by PROVEAN's algorithm or as carrying a 'high impact' by VEP, or if they were reported in the PharmGKB [35] database as having an association with any drug's metabolism, response or toxicity and constructed a support vector regression model using a radial basis function kernel to predict the AUC.

The support vector method is a universal tool for solving multidimensional nonlinear function estimation problems. In this type of algorithms the data, which are nonlinearly related to the predictor variable in the original input space, are mapped onto a higher dimensional space where a multivariate linear regression function can be constructed by means of a mapping function (known as kernel function) [36].

The regression function is such that it deviates less than a determined value of tolerance from all the training data, but only data points with deviations equal to the limits of this tolerance (known as support vectors) contribute to the model. This kind of algorithms is especially useful for high dimensional input spaces as the model does not depend on the number of dimensions of the original input space but only upon a small subset of the data (the support vectors) [37].

Support vector approaches are ideal for the analysis of biological systems thanks to their capability to solve pattern recognition problems involving small sample sizes, nonlinear relation to the predictor variable and high dimensional input data [38].

Values of the area under the atorvastatin concentration curve were natural logarithm transformed and used as target values. The number of alternative alleles for each of the 157 variants – with predicted or reported functional effects – present in atorvastatin-related genes were used to construct a support vector regression model using a radial basis function as kernel.

Model hyper parameters C (which controls complexity of the model by allowing deviations larger than the tolerance) and gamma (which controls the magnitude of influence that each data point has on the model) should be optimized to avoid overfitting. For parameter selection during model construction, an estimate of the performance can be obtained by k-fold cross validation. In this procedure, instead of dividing the initial data set into separate training and testing data sets (drastically reducing the number of samples which can be used for training), the initial data set is split into k smaller sets; one of these sets is left out for use as a validation set and the remaining sets are used for training the model, iterating until all sets have been left out in turn and the performance measure reported is then the average of all the values computed. By applying this approach we were able to optimize model hyper-parameters in such a way that encourages the highest possible reproducibility in unseen data even if the number of initial samples is limited.

Model hyper-parameters C and gamma were selected using a grid search approach. Briefly, nine equidistant values ranging from 10^{-4} to 10^{4} for each hyper-parameter were selected and a model was constructed using each combination of values while performing a threefold cross validation and the combination with a suitable correlation coefficient value was selected. Further gamma value optimization was carried out by plotting mean squared error values for the training and validation sets and selecting a gamma value which resulted in relatively low training and test sets mean squared errors and higher reproducibility in the test set (by means of a lower standard deviation of test set mean squared error).

As a next step, we calculated each variant's importance in the original model with the intention to evaluate if it was necessary to relay in all variants with predicted functional effects in atorvastatin-related genes or if a small subset of these was enough to achieve a good prediction. To this end we constructed additional regression models leaving out one of the variants in turn. We then calculated the loss in the explained vari-

Table 1. Total number of variants in 20 genes related to atorvastatin across all samples.

Variants in genes related to atorvastatin		n (%)
No predicted or reported functional effect	Known	837 (72.09)
	Novel	167 (14.38)
With a predicted or reported functional effect	Known	141 (12.14)
	Novel	16 (1.39)
Total		1161 (100)

ance of these reduced models with respect to that of the complete original model, thus allowing us to evaluate the contribution of each variant to the original model. Using this information we refined the regression model by selecting only the most important variants and evaluating how many variants were necessary to attain the maximum explained variance. Construction of all support vector regression models and internal validation analysis were carried out in Python using scikit-learn [39] machine learning tools.

Results

Target enrichment & next-generation sequencing performance

On average 3.8 (IR: 3.3–4.2) million reads per sample passed quality control. Considering all samples, 70.66% (IR: 65.72–72.99%) of target bases were covered with a depth of at least 20× and the mean target coverage depth obtained was 74.06× (IR: 64.85–83.60×).

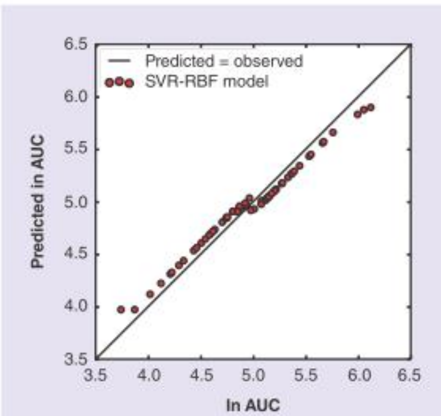


Figure 1. Original complete support vector regression model. Natural logarithm of AUC values observed and those predicted by the complete SVR-RBF. A line represents a perfect correlation between predicted and observed values.

AUC: Area under the atorvastatin concentration curve; SVR-RBF: Support Vector Regression model using a Radial Basis Function as kernel.

Variant counts

A total of 1161 variants were called across all samples in the 20 genes related to atorvastatin response and metabolism (Table 1). From the total number of variants 15.76% (183 variants) were novel and 84.24% (978 variants) were previously reported in dbSNP (version 144) [40].

Functional effect prediction

Variants in atorvastatin-related genes were considered as having a functional effect if they were reported in the PharmGKB database as associated to any drug's metabolism, response or toxicity or if they were classified as 'deleterious' by PROVEAN or as having a 'high impact' according to VEP. We considered these variants to have a possible deleterious effect on gene product activity or stability. We found a total of 157 different variants with reported or predicted functional effects across the 20 genes related to atorvastatin.

Support vector regression model

Values of the AUC were natural logarithm-transformed and used as target values. The number of alternative alleles for each of the 157 variants – with predicted or reported functional effects – present in atorvastatin-related genes were used to construct a support vector regression model using a radial basis function (hyperparameters values were C = 1 and gamma = 0.01526). This model explained 95.75% of the variance (Figure 1).

Refined regression model

We constructed 157 additional reduced support vector regression models, using information from 156 variants and leaving one variant out of each model in turn. We calculated the loss in explained variance for each reduced model in comparison to the complete model, thus allowing us to evaluate the impact of each variant in the complete model (Figure 2). We then constructed models starting with the most important variant and subsequently adding variants in order of importance from 1 to 157, with the objective of evaluating the minimum number of variants

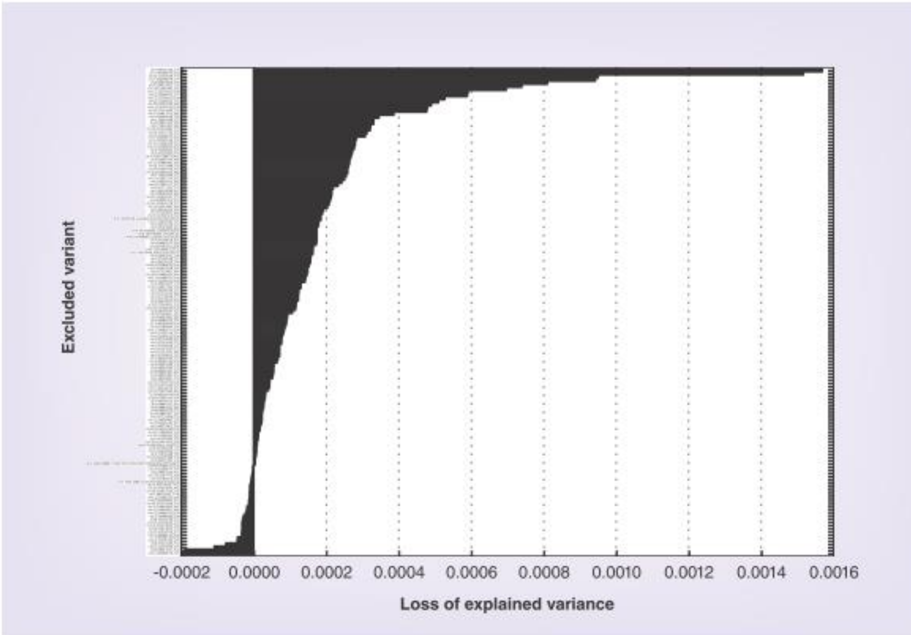


Figure 2. Importance of each variant in the complete regression model. Loss of explained variance of the reduced models with respect to the complete model as an indicator of the importance of each variant in the original complete model.

required by the model to explain a sufficient portion of the variance. This analysis showed that the maximum explained variance attainable with a model constructed using a subset of the variants was greater than that of the complete model (reaching a maximum of 95.84% of variance explained with a model constructed with 118 variants). It also showed that a considerable proportion of the variance (94.52%) could be explained with a model constructed using only the 60 most important variants (Figure 3). Importantly, 10% (six variants) of the 60 variants included in this final refined model were novel and would not have been found by other genotyping techniques.

Internal validation

In order to assess the validity of the refined regression model, we randomly selected 60 variants from the set of variants with predicted functional effects present in genes directly related to atorvastatin metabolism and response and constructed a support vector regression using those variants, iterating to 10,000,000 null hypothesis models, revealing that the p-value of obtaining a model with explained variance equal or higher than the refined model is $p < 1 \times 10^{-7}$.

Discussion

The refined support vector regression model that uses only the 60 most important variants shows a good correlation with the observed AUC values ($R^2 = 0.8979$). However, the four individuals with greater observed AUC values are clearly the less reliably represented by this model (Figure 4). Careful further analysis of these individuals might reveal other previously unknown factors, both genetic and otherwise, that modify atorvastatin pharmacokinetics.

The final regression model includes variants with a previously described important effect over atorvastatin pharmacokinetics; for example, the C allele of variant rs4149056 in *SLCO1B1* which has been associated with a 144% increase in atorvastatin AUC in homozygote individuals [41]). However, no single variant manages to explain the majority of atorvastatin pharmacokinetics in the observed individuals, and to do so, it is necessary to analyze the presence of genetic variants across several genes related to all phases of atorvastatin pharmacokinetics.

The definition of a group of genes involved in atorvastatin metabolism and response to be analyzed has a certain degree of arbitrariness and is restricted to genes included in the original sequencing experiment,

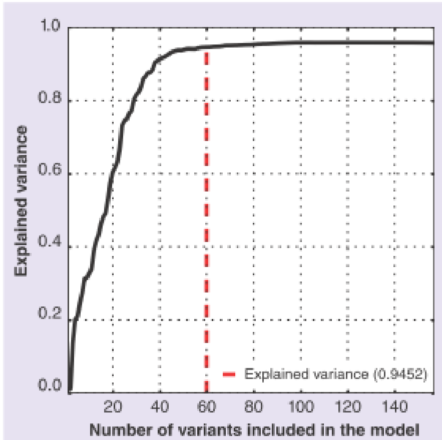


Figure 3. Explained variance of reduced models with different number of variants. Explained variance of models constructed byadding the most important variants one by one until the total (175 variants) is reached. A dotted line highlights that a substantial portion of the explained variance (94.52%) is attained with the construction of a reduced model using the 60 most important variants.

however, preliminary results of this work support the idea that including a greater number of genes with reported associations to a drug's pharmacokinetics or pharmacodynamics profiles improves the performance of the model in terms of the proportion of the variance explained.

According to our analysis, in order to predict atorvastatin pharmacokinetics parameters (AUC), an integrated analysis of several genes known to intervene in the absorption, Phase I and II metabolism, excretion and response to this drug is needed (Table 2) and no single variant determines the pharmacokinetic profile of atorvastatin on its own.

It is generally accepted that the directing steps in atorvastatin pharmacokinetics are the transporters OATP1B1 (encoded in *SLCO1B1*), P-glycoprotein (encoded in *ABCB1*) or BCRP (encoded in *ABCG2*). This is partially supported by our findings, as the greater number of variants included in our final refined model were found in genes involved in atorvastatin transport (*SLCO1B1* and *ABCB1* which presented eight and six variants, respectively). However, the two variants with greater importance to the model are located in the *CYP3A4* gene, which is involved in the Phase I metabolism of atorvastatin, and highlights the necessity of including the different phases of atorvastatin's metabolism and response in the analysis.

The use of a group of genetic variants prioritized by relative importance to construct a model explaining the variability in the AUC of atorvastatin presented in this work differs significantly from other recent pharmacogenetic studies using next-generation sequencing [42,43] as they aimed to identify single variants presenting a substantial effect on the observed phenotype. Nevertheless, all of these studies highlight the paramount importance of capturing both common and rare genetic variants, either novel or known in order to explain the interpersonal variability in drug response.

The fact that, in our case, no single variant was capable of explaining the majority of atorvastatin AUC may be due to the fact that some of the variants previously reported as fundamental to variation in statin metabolism were found only in relatively low frequencies, and thus other variants are needed to explain the interpersonal variation observed in atorvastatin pharmacokinetics in this population sample (Table 2). For example, rs2231142, which produces a Gln141Lys change in *ABCG2'*s gene product and which is associated with higher plasma concentrations of atorvastatin is present with a frequency of 16.67% and rs4149056 which produces a Val174Ala change in the transporter encoded in *SLCO1B1*, which is also associated with a higher plasma concentration of atorvastatin, was present with a frequency of only 10.83%.

The inverse, where variants are so frequent that additional variants may be needed to explain the observed differences in atorvastatin pharmacokinetics, is also true, as other variants that have also been consistently considered of pharmacogenetic importance

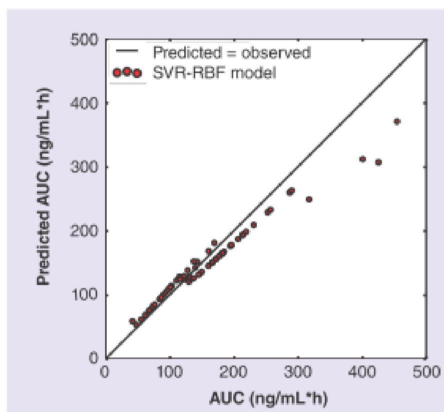


Figure 4. Refined support vector regression model. AUC values observed and predicted by the final refined SVR-RBF. A line represents a perfect correlation between predicted and observed values.
AUC: Area under the atorvastatin concentration curve; SVR-RBF: Support Vector Regression model using a Radial Basis Function as kernel.

Table 2. Variants included in the final regression model.

Importance to the model in descending order	Variant location and alleles	Gene	dbSNP (version 144)	Study AF (n = 60)[†]
1	chr7:99361466 C/T	*CYP3A4*	rs2242480	0.3917
2	chr7:99360870 G/A	*CYP3A4*	rs4646440	0.2167
3	chr12:21382619 T/C	*SLCO1B1*	rs11045879	0.1417
4	chr7:150690079 C/T	*NOS3*	rs2070744	0.7167
5	chr4:89078924 T/C	*ABCG2*	rs2622604	0.7750
6	chr2:234637022 T/C	*UGT1A4/UGT1A3*	rs1983023	0.1917
7	chr17:61565990 G/C	*ACE*	rs4341	0.4417
8	chr12:21011581 G/A	*SLCO1B3*	rs4149118	0.6250
9	chr12:20984832 C/T	*SLCO1B3*	rs10841661	0.3833
10	chr12:21378021 G/A	*SLCO1B1*	rs4149081	0.1083
11	chr12:21368722 T/C	*SLCO1B1*	rs4363657	0.1333
12	chr12:21329738 A/G	*SLCO1B1*	rs2306283	0.4250
13	chr12:21015760 G/A	*SLCO1B3*	rs7311358	0.7833
14	chr12:21331549 T/C	*SLCO1B1*	rs4149056	0.1083
15	chr2:234637015 A/G	*UGT1A4/UGT1A3*	rs2008584	0.3333
16	chr2:234665659 T/G	*UGT1A4/UGT1A3*	rs4124874	0.4917
17	chr7:75615369 C/T	*POR*	rs72557950	0.0250
18	chr4:69972949 C/G	*UGT2B7*	rs4292394	0.0667
19	chr5:74651084 A/G	*HMGCR*	rs3846662	0.4333
20	chr10:101591944 T/C	*ABCC2*	rs4148396	0.6500
21	chr12:21317791 C/T	*SLCO1B1*	rs4149032	0.3583
22	chr12:21331625 C/T	*SLCO1B1*	rs2291075	0.2833
23	chr7:150696111 T/G	*NOS3*	rs1799983	0.7833
24	chr17:61583756 C/T	*ACE*	rs4267385	0.3417
25	chr12:21028208 G/C	*SLCO1B3*	rs60140950	0.1167
26	chr22:42526694 G/A	*CYP2D6*	rs1065852	0.0583
27	chr5:74648603 A/T	*HMGCR*	rs12654264	0.3833
28	chr7:87229440 T/C	*ABCB1*	rs9282564	0.0583
29	chr7:150713902 G/GC	*NOS3*	rs11393607	0.1667
30	chr7:99365719 A/G	*CYP3A4*	rs2246709	0.4750
31	chr12:20980800 G/C	*SLCO1B3*	rs7977213	0.6833
32	chr12:21331599 T/C	*SLCO1B1*	rs4149057	0.5833
33	chr17:61584720 A/T	*ACE*	rs4459610	0.3500
34	chr10:101541053 A/G	*ABCC2*	rs1885301	0.6167
35	chr7:75615006 C/T	*POR*	rs1057868	0.3750
36	chr22:42522724 G/T	*CYP2D6*	rs79392742	0.0250
37	chr1:66058513 A/G	*LEPR*	rs1137101	0.4667
38	chr9:107645477 G/T	*ABCA1*	rs12003906	0.0417

[†]Allelic frequencies of the called variant allele, reference allele obtained from the human genome build hg19.
AF: Allelic frequency.

Table 2. Variants included in the final regression model (cont.).				
Importance to the model in descending order	Variant location and alleles	Gene	dbSNP (version 144)	Study AF (n = 60)[†]
39	chr7:87171152 T/C	*ABCB1*	rs4148737	0.3000
40	chr7:99245080 A/C	*CYP3A5*	rs4646457	0.2000
41	chr7:99245914 A/G	*CYP3A5*	rs15524	0.2000
42	chr10:101604207 C/T	*ABCC2*	rs3740066	0.3167
43	chr4:89052323 G/T	*ABCG2*	rs2231142	0.1667
44	chr10:101542578 C/T	*ABCC2*	rs717620	0.1083
45	chr22:42523943 A/G	*CYP2D6*	rs16947	0.7583
46	chr7:99245013 T/G	*CYP3A5*	rs4646458	0.1000
47	chr7:87138645 A/G	*ABCB1*	rs1045642	0.5000
48	chr22:42522613 G/C	*CYP2D6*	rs1135840	0.6583
49	chr5:74638526 G/GAGACCGCACCAGCGACAC	*HMGCR*		0.0333
50	chr11:74907582 C/T	*SLCO2B1*	rs2306168	0.1250
51	chr7:99270539 C/T	*CYP3A5*	rs776746	0.2083
52	chr7:87164986 A/G	*ABCB1*	rs10248420	0.2250
53	chr9:107599265 A/AGACCGCAC	*ABCA1*		0.0083
54	chr9:107599265 A/AGACCG	*ABCA1*		0.0167
55	chr9:107599265 A/AGACCGCACCAG	*ABCA1*		0.0333
56	chr7:87160618 A/C	*ABCB1*	rs2032582	0.5083
57	chr4:69964271 A/G	*UGT2B7*	rs28365062	0.3583
58	chr7:87157051 G/A	*ABCB1*	rs7787082	0.2167
59	chr7:75614946 C/CCG	*POR*		0.0250
60	chr7:75614948 A/ACCAGCGACAC	*POR*		0.0250

[†]Allelic frequencies of the called variant allele, reference allele obtained from the human genome build hg19.
AF: Allelic frequency.

were much more frequent, to the point of being present in almost all individuals. For example, rs2740574, a -392 A>G change in the 5′-flanking region of *CYP3A4* which increases transcription of a cytochrome P450 enzyme involved in atorvastatin's Phase I metabolism was found in our population sample with a frequency of 92.5%. In this regard, it is important to note that the 60 variants included in the final regression model had allelic frequencies with median of 29.17% (IR: 10.83–46.88%), although nine variants presented frequencies lower than 5% and can thus be considered rare in this population sample.

In spite of a relatively small sample size, our results suggest that the majority of the variation in atorvastatin pharmacokinetics can be explained using a number of genetic markers supported by the prediction of the functional effect of each variant. However, an evaluation of the regression model in an independent sample is the only means to obtaining an accurate validation of the model's performance in new individuals.

Currently, it is impossible to evaluate the extent to which this model can be applied to populations with a genetic background significantly different than the one used during the construction of the model (in this case healthy northeastern Mexicans), and to achieve this goal extensive evaluation in such population samples would be needed.

As seen in all studies trying to link genetic variation to a specific phenotype, some of the variants included in the regression model may not be the ones responsible for the effect on atorvastatin pharmacokinetics. Consequently, functional studies are required in order to ascertain whether these variants have an actual effect or if they are markers closely related to other underlying functional variants, even if the use of informatic prediction tools should diminish the possibility of such events.

An important barrier for the implementation of the comprehensive analysis of many genes relevant not only to atorvastatin but also for a wide range of other drugs' metabolism and response in current clinical practice is the high cost associated with next-generation sequencing experiments for discovery. However, the dropping costs of genotyping technologies, the increasing number of validated pharmacogenetic markers and the now increasing body of pre-emptively genotypified individuals [44] contribute to the usefulness of approaches similar to the one used on this work, and their future implementation in clinical practice.

Conclusion

In this study we used genetic variants with predicted or reported functional effects present across 20 genes to explain the majority of the variance in atorvastatin AUC. These findings support the use of comprehensive approaches where several genes involved in different phases of drug metabolism and response are analyzed simultaneously to explain to a greater extent a specific phenotype, in this case atorvastatin concentration profiles of 60 healthy volunteers. Nevertheless, further studies are required in order to assess the usefulness and clinical relevance of this model in independent population samples, specially in those with a significantly different genetic background.

Acknowledgements

The authors would like to thank the members of INMEGEN's sequencing unit, specially Julio César Canseco for expert assistance in next-generation library preparation and sequencing and Everardo Piñeyro Garza for his support in volunteer recruitment and pharmacokinetic characterization.

Financial & competing interests disclosure

OF Cruz-Correa is a doctoral student from the Doctorate Program in Biochemical Sciences of the Universidad Nacional Autónoma de México (UNAM) and received fellowship 366725/245614 from the CONACYT. Research was supported by CONACyT's Convocatoria de Investigación Científica Básica (grant #252952 "Diversidad Fármacogenética en Mexicanos, colección e interpretación"), Fondo del laboratorio de Ciencia y Tecnología (grant #124140 "Servicios Especializados de Investigación, Desarrollo e Innovación para Farmoquímicos y Biotecnológicos"), Programa de Estímulo a la Investigación, Desarrollo Tecnológico e Innovación (#218098 "Farmacogenética de la Diabetes Mellitus tipo 2") and Fondo de Innovación Tecnológica (grant #260826 "Farmacogenética y Medicina de Precisión"). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Executive summary

- Pharmacogenetics has the potential to attain higher drug efficacy and lower adverse reactions through the personalization of drug therapy regimes.
- In most cases, genetic variants known to modify a drug therapy regime fail to completely explain the interpersonal variability in drug pharmacokinetics.
- Genetic variants, sometimes specific to the studied population, might have an effect on different steps of metabolism, transport and excretion of the same drug.
- Initially, we used variants with a predicted reported functional effect in 20 genes known to be involved in atorvastatin metabolism and response to construct a regression model of atorvastatin's area under the plasmatic concentration curve values.
- In order to explain a greater extent of the variance, we then refined the model by choosing a smaller set of variants which contributed the most to the model.
- The final support vector regression model using 60 variants explained a substantial proportion of the variance (94.52%) in atorvastatin area under the plasma concentration curve.
- The results highlight the need of an integrated analysis of several genes known to intervene in the different steps of metabolism of atorvastatin in order to predict atorvastatin area under the plasmatic concentration curve values.
- An approach similar to the one used here would probably be required not only for atorvastatin but also for a wide range of other drugs.
- However, extensive validation of the model in populations with different genetic backgrounds and the evaluation of each variant's effects through functional studies are still required.

References

Papers of special note have been highlighted as: • of interest

1 León-Cachón RB, Ascacio-Martínez JÁ, Gómez-Silva M *et al.* Application of genomic technologies in clinical pharmacology research. *Rev. Invest. Clin.* 67(4), 212–218 (2015).

2 Clinical Pharmacogenetics Implementation Consortium. CPIC genes–drugs (2016). https://cpicpgx.org/genes-drugs/

• **List of peer-reviewed dosing guidelines based on genetic information.**

3 Hernandez W, Gamazon ER, Aquino-Michaels K *et al.* Ethnicity-specific pharmacogenetics: the case of warfarin in African Americans. *Pharmacogenomics J.* 14(3), 223–228 (2014).

• **Study showing the poorer performance of standardized pharmacogenetic dosing algorithms in certain populations and that considering additional genetic variants relevant to such populations help improve on these algorithms.**

4 Fernandez G, Spatz ES, Jablecki C, Phillips PS. Statin myopathy: a common dilemma not reflected in clinical trials. *Cleve. Clin. J. Med.* 78(6), 393–403 (2011).

5 Lennernäs H. Clinical pharmacokinetics of atorvastatin. *Clin. Pharmacokinet.* 42(13), 114–160 (2003).

• **Nonrecent but very complete review of atorvastatin basic pharmacokinetics.**

6 Chen C, Mireles RJ, Campbell SD *et al.* Differential interaction of 3-hydroxy-3-methylglutaryl-COA reductase inhibitors with ABCB1, ABCC2, and OATP1B1. *Drug Metab. Dispos.* 33(4), 537–546 (2005).

7 Karlgren M, Vildhede A, Norinder U *et al.* Classification of inhibitors of hepatic organic anion transporting polypeptides (OATPs): influence of protein expression on drug–drug interactions. *J. Med. Chem.* 55(10), 4 (2012).

8 Park JE, Kim KB, Bae SK, Moon BS, Liu KH, Shin JG. Contribution of cytochrome P450 3A4 and 3A5 to the metabolism of atorvastatin. *Xenobiotica* 38, 1240–1251 (2008).

9 Goosen TC, Bauman JN, Davis JA *et al.* Atorvastatin glucuronidation is minimally and nonselectively inhibited by the fibrates gemfibrozil, fenofibrate, and fenofibric acid. *Drug Metab. Dispos.* 35(8), 1315–1324 (2007).

10 Prueksaritanont T, Subramanian R, Fang X *et al.* Glucuronidation of statins in animals and human: a novel mechanism of statin lactonization. *Drug Metab. Dispos.* 30(5), 505–512 (2002).

11 Voora D, Shah SH, Reed CR *et al.* Pharmacogenetic predictors of statin-mediated low-density lipoprotein cholesterol reduction and dose response. *Circ. Cardiovasc. Genet.* 1(2), 100–106 (2008).

12 Potaczek DP, Undas A, Iwaniec T, Szczeklik A. The angiotensin-converting enzyme gene insertion/deletion polymorphism and effects of quinapril and atorvastatin on haemostatic parameters in patients with coronary artery disease. *Thromb. Haemost.* 94(1), 224–225 (2009).

13 Donnelly LA, Doney AS, Dannfald J *et al.* A paucimorphic variant in the HMG-CoA reductase gene is associated with lipid-lowering response to statin treatment in diabetes: a GoDARTS study. *Pharmacogenet. Genomics* 18(12), 1021–1026 (2008).

14 Chien KL, Wang KC, Chen YC *et al.* Common sequence variants in pharmacodynamic and pharmacokinetic pathway-related genes conferring LDL cholesterol response to statins. *Pharmacogenomics* 11(3), 309–317 (2010).

15 Chung JY, Cho SK, Oh ES *et al.* Effect of HMGCR variant alleles on low-density lipoprotein cholesterol-lowering response to atorvastatin in healthy Korean subjects. *J. Clin. Pharmacol.* 52(3), 339–346 (2012).

16 Cuevas A, Fernández C, Ferrada L *et al. HMGCR* rs17671591 SNP determines lower plasma LDL-C after atorvastatin therapy in Chilean individuals. *Basic Clin. Pharmacol. Toxicol.* 118(4), 292–297 (2016).

17 Drogari E, Ragia G, Mollaki V, Elens L, Van Schaik RH, Manolopoulos VG. *POR*28* SNP is associated with lipid response to atorvastatin in children and adolescents with familial hypercholesterolemia. *Pharmacogenomics* 15(16), 1963–1972 (2014).

18 Cho SK, Oh ES, Park K, Park MS, Chung JY. The *UGT1A3*2* polymorphism affects atorvastatin lactonization and lipid-lowering effect in healthy volunteers. *Pharmacogenet. Genomics* 22(8), 598–605 (2012).

19 Willrich MA, Hirata MH, Genvigir FD *et al. CYP3A53A* allele is associated with reduced lowering-lipid response to atorvastatin in individuals with hypercholesterolemia. *Clin. Chim. Acta.* 398(1–2), 15–20 (2008).

20 Becker ML, Visser LE, van Schaik RHN, Hofman A, Uitterlinden AG, Stricker BHC. Influence of genetic variation in *CYP3A4* and *ABCB1* on dose decrease or switching during simvastatin and atorvastatin therapy. *Pharmacoepidemiol. Drug Safet.* 19(1), 75–81 (2010).

21 Ulvestad M, Skottheim IB, Jakobsen GS *et al.* Impact of *OATP1B1, MDR1*, and *CYP3A4* expression in liver and intestine on interpatient pharmacokinetic variability of atorvastatin in obese subjects. *Clin. Pharmacol. Ther.* 93(3), 275–282 (2013).

22 Keskitalo JE, Zolk O, Fromm MF, Kurkinen KJ, Neuvonen PJ, Niemi M. *ABCG2* polymorphism markedly affects the pharmacokinetics of atorvastatin and rosuvastatin. *Clin. Pharmacol. Ther.* 86(2), 197–203 (2009).

23 Becker ML, Elens LL, Visser LE *et al.* Genetic variation in the *ABCC2* gene is associated with dose decreases or switches to other cholesterol-lowering drugs during simvastatin and atorvastatin therapy. *Pharmacogenomics J.* 13(3), 251–256 (2013).

24 Frudakis TN, Thomas MJ, Ginjupalli SN, Handelin B, Gabriel R, Gomez HJ. *CYP2D6*4* polymorphism is associated with statin-induced muscle effects. *Pharmacogenet. Genomics* 17(9), 695–707 (2007).

25 Pérez-Castrillón JL, Vega G, Abad L *et al.* Atorvastatin and BMD in coronary syndrome. Role of Lys656Asn polymorphism of leptin receptor gene. *Endocr. J.* 56(2), 221–225 (2009).

26 Pérez-Castrillón JL, Vega G, Abad L *et al.* Effect of the TNFalpha-308 G/A polymorphism on the changes produced by atorvastatin in bone mineral density in patients with

acute coronary syndrome. *Ann. Nutr. Metab.* 53(2), 117–121 (2008).

27 Leon-Cachon RB, Ascacio-Martinez JA, Gamino-Peña ME *et al.* A pharmacogenetic pilot study reveals *MTHFR, DRD3* and *MDR1* polymorphisms as biomarker candidates for slow atorvastatin metabolizers. *BMC Cancer* 16, 74 (2016).

28 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), 2114–2120 (2014).

29 Broad Institute. https://github.com/broadinstitute/picard

30 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14), 1754–1760 (2009).

31 McKenna A, Hanna M, Banks E *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9), 1297–1303 (2010).

32 Van der Auwera GA, Carneiro M, Hartl C *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 11(1110), 11.10.1–11.10.33 (2013).

33 Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 7(10), e46688 (2012).

34 McLaren W, Pritchard B, Rios D, Chen Y, Flicek P Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26(16), 2069–2070 (2010).

35 Whirl-Carrillo M, McDonagh EM, Hebert JM *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* 92(4), 414–417 (2012).

• **Pharmacogenomics Knowledge base (PharmGKB) collects and curates information about the impact of human genetic variation on drug response.**

36 Vapnik V, Golowich SE, Smola A. Support vector for function approximation, regression estimation and signal processing. Advances in Neural Information Processing Systems 9. MIT Press, Cambridge, MA, USA, 281–287 (1997).

37 Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge, UK (2000).

38 Ustün B, Melssen WJ, Buydens LM. Visualisation and interpretation of support vector regression models. *Anal. Chim. Acta.* 595(1–2), 299–309 (2007).

39 Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12(Oct), 2825–2830 (2011).

40 Sherry ST, Ward MH, Kholodov M *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1), 308–311 (2001).

41 Pasanen MK, Fredrikson H, Neuvonen PJ, Niemi M. Different effects of *SLCO1B1* polymorphism on the pharmacokinetics of atorvastatin and rosuvastatin. *Clin. Pharmacol. Ther.* 82(6), 726–733 (2007).

42 Mizzi C, Peters B, Mitropoulou C *et al.* Personalized pharmacogenomics profiling using whole-genome sequencing. *Pharmacogenomics* 15(9), 1223–1234 (2014).

43 Tavira B, Gómez J, Diaz-Corte C *et al.* *ABCB1* (*MDR-1*) pharmacogenetics of tacrolimus in renal transplanted patients: a Next Generation Sequencing approach. *Clin Chem Lab Med.* 53(10), 1515–1519 (2015).

44 Hoffman JM, Haidar CE, Wilkinson MR *et al.* PG4KDS: a model for the clinical implementation of pre-emptive pharmacogenetics. *Am. J. Med. Genet. C Semin Med. Genet.* 166C(1), 45–55 (2014).

Anexo 2. Listado y características de las 60 variantes incluidas en el modelo final de regresión de los perfiles de concentración plasmática (AUC) de atorvastatina.

Importancia en el modelo (descendente)	Variante	Gen	dbSNP (versión 144)	Frecuencia (Atorvastatina, n=60)	Frecuencia (Amerindios, n=94)	Frecuencia (Exomas, n=968)	Frecuencia (Kaviar, n=77,781)	Predicción (PROVEAN)	Predicción (VEP)	Predicción no codificante (Fathmm-mkl)	Predicción Codificante (Fathmm-mkl)	Asociación a fármacos (PharmGKB)
1	chr7:99360870_G/A	CYP3A4	rs4646440	0.2167	0.4628	NA	0.0854		MODIFIER	0.1592	0.0507	presente
2	chr7:99361466_C/T	CYP3A4	rs2242480	0.3917	0.6543	0.5506	0.2082		MODIFIER	0.1222	0.0035	presente
3	chr12:21382619_T/C	SLCO1B1	rs11045879	0.1417	0.1330	NA	0.2006		MODIFIER	0.0822	0.0025	presente
4	chr9:107690450_G/A	ABCA1	rs1800977	0.3879	0.5957	NA	0.3448		MODIFIER	0.9556	0.7653	ausente
5	chr4:89078924_T/C	ABCG2	rs2622604	0.7750	0.8883	NA	0.7595		MODIFIER	0.0828	0.0046	presente
6	chr17:61565990_G/C	ACE	rs4341	0.4732	0.7340	0.6508	0.4398		MODIFIER	0.0611	0.0014	presente
7	chr7:150690079_C/T	NOS3	rs2070744	0.7167	0.9628	NA	0.6814		MODIFIER	0.1310	0.0169	presente
8	chr7:150704250_C/G	NOS3	rs2566514	0.5000	0.7766	0.6958	0.3196	Neutral	MODIFIER	0.9171	0.0246	ausente
9	chr12:21378021_G/A	SLCO1B1	rs4149081	0.1083	0.1170	NA	0.1983		MODIFIER	0.1048	0.0143	presente
10	chr4:89061114_C/T	ABCG2	rs2231137	0.2250	0.3511	0.3027	0.0962	Neutral	MODERATE	0.7702	0.0009	ausente
11	chr2:234637022_T/C	UGT1A4/ UGT1A3	rs1983023	0.2396	0.3511	NA	0.3705		MODIFIER	0.0575	0.0027	presente
12	chr12:21011581_G/A	SLCO1B3	rs4149118	0.6250	0.6011	NA	0.6787		MODIFIER	0.1399	0.0213	presente
13	chr12:21329738_A/G	SLCO1B1	rs2306283	0.4250	0.4149	0.4189	0.4545	Neutral	MODERATE	0.0890	0.0118	presente
14	chr12:21011480_T/G	SLCO1B3	rs4149117	0.7917	0.7128	0.7800	0.7762	Neutral	MODERATE	0.9540	0.0961	ausente
15	chr12:21368722_T/C	SLCO1B1	rs4363657	0.1333	0.1330	NA	0.1991		MODIFIER	0.0452	0.0011	presente
16	chr9:107588033_C/T	ABCA1	rs2066715	0.0750	0.0160	0.0527	0.0790	Neutral	MODERATE	0.9686	0.7715	ausente
17	chr7:75615006_C/T	POR	rs1057868	0.3814	0.3138	0.2893	0.2526	Neutral	MODERATE	0.2711	0.1120	presente
18	chr12:20984832_C/T	SLCO1B3	rs10841661	0.3833	0.2660	NA	0.4086		MODIFIER	0.0320	0.0001	presente
19	chr17:61566031_G/A	ACE	rs4343	0.6333	0.7340	0.6658	0.4823	Neutral	LOW	0.1034	0.0179	presente
20	chr2:234627937_T/C	UGT1A4	rs2011404	0.8917	0.9574	NA	0.7696	Neutral	MODIFIER	0.5891	0.0973	ausente
21	chr12:21331549_T/C	SLCO1B1	rs4149056	0.1083	0.1170	0.0904	0.1258	Deleterious	MODERATE	0.7254	0.9809	presente
22	chr12:21015760_G/A	SLCO1B3	rs7311358	0.7833	0.7074	NA	0.7763	Neutral	MODERATE	0.7554	0.1128	presente
23	chr4:89052323_G/T	ABCG2	rs2231142	0.1667	0.4149	0.2903	0.1150	Neutral	MODERATE	0.9823	0.8349	presente
24	chr4:69972949_C/G	UGT2B7	rs4292394	0.2500	0.8670	NA	0.5169	Neutral	LOW	0.0564	0.0071	presente
25	chr9:107665751_C/T	ABCA1	rs2482424	0.0816	0.3670	NA	0.1420		MODIFIER	0.5370	0.1102	ausente
26	chr9:107591272_G/T	ABCA1	rs2853579	0.2373	0.1809	0.2133	0.1928	Neutral	LOW	0.9910	0.9602	ausente
27	chr17:61573761_T/C	ACE	rs4362	0.6271	0.7340	0.6555	0.4978	Neutral	MODIFIER	0.7566	0.1081	ausente
28	chr22:42526694_G/A	CYP2D6	rs1065852	0.0729	0.1237	0.0957	0.1763	Deleterious	MODERATE	0.9890	0.9850	presente

29	chr12:21331625_C/T	SLCO1B1	rs2291075	0.2833	0.1344	0.2025	0.3701	Neutral	LOW	0.8992	0.0583	presente
30	chr9:107620835_G/A	ABCA1	rs9282541	0.1167	0.1543	0.1162	0.0081	Neutral	MODIFIER	0.4055	0.7596	ausente
31	chr10:101591944_T/C	ABCC2	rs4148396	0.6500	0.6755	NA	0.7396		MODIFIER	0.1519	0.0033	presente
32	chr17:61585072_G/A	ACE	rs112549572	0.0167	NA	NA	0.0008		MODIFIER	0.8295	0.6418	ausente
33	chr7:75544455_A/C	POR	rs3823884	0.2966	0.5426	NA	0.3421		MODIFIER	0.8592	0.7334	ausente
34	chr7:150713902_G/GC	NOS3	rs76257575	1.0000	1.0000	NA	0.0290		HIGH	NA	NA	ausente
35	chr5:74648603_A/T	HMGCR	rs12654264	0.3833	0.4202	NA	0.4106		MODIFIER	0.0332	0.0001	presente
36	chr7:75615369_C/T	POR	rs72557950	0.0250	NA	NA	0.0001	Deleterious	MODERATE	0.8746	0.6147	ausente
37	chr5:74651084_A/G	HMGCR	rs3846662	0.4333	0.4202	0.4545	0.4747		MODIFIER	0.1367	0.0016	presente
38	chr17:61559923_C/T	ACE	rs4309	0.5932	0.7660	0.6880	0.4310	Neutral	MODIFIER	0.5086	0.0616	ausente
39	chr12:21317791_C/T	SLCO1B1	rs4149032	0.3583	0.3138	0.3401	0.3038		MODIFIER	0.0459	0.0002	presente
40	chr12:21028208_G/C	SLCO1B3	rs60140950	0.1167	0.0053	0.0460	0.1051	Deleterious	MODERATE	0.9855	0.9623	ausente
41	chr17:61583756_C/T	ACE	rs4267385	0.3417	0.1915	NA	0.5069		MODIFIER	0.0400	0.0001	presente
42	chr9:107602666_C/T	ABCA1	rs2246841	0.1356	0.2447	0.1839	0.1158	Neutral	LOW	0.9332	0.0996	ausente
43	chr4:89015857_C/T	ABCG2	rs2231164	0.5583	0.2234	0.4236	0.7176		MODIFIER	0.4459	0.1128	presente
44	chr7:87229440_T/C	ABCB1	rs9282564	0.0583	NA	0.0196	0.0738	Neutral	MODERATE	0.2765	0.0469	presente
45	chr17:61586549_T/C	ACE	rs10853044	0.3417	0.1915	NA	0.4684		MODIFIER	0.9920	0.9862	ausente
46	chr22:42522724_G/T	CYP2D6	rs79392742	0.0259	0.0806	NA	0.0020	Deleterious	MODERATE	0.9942	0.9974	ausente
47	chr4:69964271_A/G	UGT2B7	rs28365062	0.3644	0.5167	0.4711	0.1500	Neutral	LOW	0.2682	0.0716	presente
48	chr2:234665659_T/G	UGT1A4/ UGT1A3	rs4124874	0.4917	0.5479	NA	0.4671		MODIFIER	0.9272	0.0011	presente
49	chr2:234622429_T/C	UGT1A5	rs2012734	0.5000	0.5638	NA	0.4661	Neutral	MODIFIER	0.5631	0.0131	ausente
50	chr17:61584720_A/T	ACE	rs4459610	0.3500	0.1915	NA	0.0925		MODERATE	0.1510	0.0206	presente
51	chr11:74907582_C/T	SLCO2B1	rs2306168	0.1293	0.1223	0.0997	0.0750	Neutral	MODERATE	0.6657	0.0275	presente
52	chr7:150696111_T/G	NOS3	rs1799983	0.7833	0.9096	0.8704	0.7116	Neutral	MODERATE	0.9274	0.0947	presente
53	chr5:74638526_G/GAGACCGCACCAGCGACAC	HMGCR		0.0333	NA	NA	NA	Deleterious	MODERATE	NA	NA	ausente
54	chr12:21331599_T/C	SLCO1B1	rs4149057	0.5833	0.4894	0.5289	0.5119	Neutral	LOW	0.8913	0.0138	presente
55	chr10:101542578_C/T	ABCC2	rs717620	0.1083	0.0904	0.1018	0.1665		MODIFIER	0.2651	0.3006	presente
56	chr10:101541053_A/G	ABCC2	rs1885301	0.6167	0.6436	NA	0.6019		MODIFIER	0.1257	0.0111	presente
57	chr7:99270539_C/T	CYP3A5	rs776746	0.2083	0.3032	NA	0.1823		HIGH	0.0378	0.0002	presente
58	chr11:74883577_G/A	SLCO2B1	rs12422149	0.3707	0.7340	0.5511	0.1620	Neutral	MODERATE	0.1715	0.0225	presente
59	chr11:74907721_C/T	SLCO2B1	rs61555831	0.0333	0.0106	0.0093	0.0307	Neutral	LOW	0.9828	0.7836	ausente
60	chr17:61588126_C/T	ACE	rs7221780	0.3583	0.1915	NA	0.3742		MODIFIER	0.9626	0.1450	ausente

NA en las columnas de frecuencia indica que la variante no se encontró presente en ese grupo de datos, ya sea porque ningún individuo es portador o porque la región donde se encuentra la variante no fue analizada (en el caso de exomas). NA para Fathmm-MKL indica que las predicciones con esta herramienta no están disponibles ya que no es un cambio de un solo nucleótido.

Anexo 3. Número de variantes por extensión del gen secuenciada para cada uno de los genes importantes de farmacogenómica incluidos en el análisis.

En negritas se resaltan los cinco valores más altos para cada una de las categorías.

Gen	Variantes/kbp	Variantes con efectos funcionales/kbp	Variantes nuevas/kbp
ABCA1	10.32	3.65	1.56
ABCA4	11.40	4.20	1.19
ABCA8	9.89	2.55	1.67
ABCA12	9.03	2.61	0.98
ABCB1	9.73	3.69	3.09
ABCB4	9.96	2.13	2.19
ABCB5	12.60	3.23	1.89
ABCB6	11.24	5.56	1.52
ABCB7	4.97	1.49	0.37
ABCB8	12.81	2.76	1.49
ABCB9	7.95	2.54	1.78
ABCB11	11.83	2.82	2.00
ABCC1	12.61	1.87	1.83
ABCC2	9.41	2.78	1.13
ABCC3	9.40	2.56	1.36
ABCC4	11.91	2.20	1.77
ABCC5	7.75	2.73	1.26
ABCC6	12.80	2.86	1.58
ABCC8	10.58	3.00	1.37
ABCC9	7.75	1.14	1.96
ABCC10	8.75	3.38	1.52
ABCC11	11.76	3.06	1.87
ABCC12	8.91	4.27	1.37
ABCD1	6.17	0.94	0.54
ABCD2	7.03	2.34	1.87
ABCG1	10.46	1.70	0.88
ABCG2	7.89	2.07	0.77
ABO	**32.90**	3.87	4.98
ACBD4	7.89	2.12	1.15
ACE	11.30	3.16	1.54
ACP2	13.24	2.47	2.82
ADA	10.48	3.60	1.47
ADAMTS1	10.35	3.45	1.04
ADCK4	11.56	3.68	0.92
ADD1	6.89	1.23	1.77

ADH1A	9.77	1.76	2.56
ADH1B	10.92	2.60	2.47
ADH1C	15.13	2.16	1.66
ADH4	10.57	2.52	1.34
ADH5	10.08	1.24	1.52
ADH6	8.65	1.82	1.82
ADH7	9.20	1.90	1.46
ADHFE1	11.51	2.50	1.50
ADRB1	16.20	5.63	**8.80**
ADRB2	17.23	8.37	0.49
AGT	15.36	5.39	1.89
AGTR1	10.35	2.24	1.38
AHR	6.82	1.78	1.68
AKAP9	8.21	2.21	1.53
ALB	12.79	3.28	4.87
ALDH1A1	7.85	1.59	2.01
ALDH1A2	10.05	1.32	1.75
ALDH1A3	8.39	1.05	1.62
ALDH1B1	10.80	4.09	0.88
ALDH2	8.67	2.49	1.31
ALDH3A1	12.48	3.44	2.04
ALDH3A2	7.38	1.49	1.58
ALDH3B1	11.83	1.84	2.14
ALDH3B2	16.68	5.20	1.38
ALDH4A1	14.84	4.73	1.40
ALDH5A1	12.30	2.03	3.00
ALDH6A1	14.89	4.23	2.51
ALDH7A1	11.13	1.81	2.41
ALDH8A1	8.77	2.15	1.49
ALDH9A1	13.56	2.56	1.51
ALOX5	12.88	3.04	0.97
AOX1	9.97	2.15	0.89
APOA2	**55.99**	7.89	**14.20**
APOE	8.75	2.59	2.59
ARNT	10.17	1.78	2.48
ARNTL	7.16	1.49	2.20
ARSA	16.76	3.39	2.26
ARVCF	17.66	5.48	2.47
ASNA1	7.14	3.89	0.65
ATF6B	7.12	2.62	1.37
ATG9B	17.82	5.65	2.97
ATP7A	11.15	6.37	**7.12**
ATP7B	10.51	3.92	0.89

BCHE	10.93	4.60	2.30
BDH2	7.14	1.05	1.58
BDKRB2	10.32	0.89	2.17
BHLHE22	8.54	5.89	1.77
BRCA1	7.65	2.11	1.08
C3orf36	17.41	1.80	4.20
CAT	8.53	2.19	1.46
CBR1	16.51	4.45	3.49
CBR3	18.79	**9.11**	1.14
CDA	13.92	2.71	0.39
CDKN1C	10.21	3.14	2.36
CES1	17.72	3.02	3.02
CES2	7.22	1.74	1.62
CETP	13.99	2.89	0.91
CFTR	9.36	3.76	1.43
CHRNA2	11.55	2.18	1.80
CHST1	8.34	2.95	1.04
CHST2	9.67	4.72	1.65
CHST3	9.90	1.78	1.78
CHST4	10.66	4.04	2.21
CHST5	11.62	2.90	3.68
CHST6	10.66	2.98	2.11
CHST7	5.12	4.02	1.46
CHST8	9.31	3.10	1.36
CHST9	6.22	1.72	1.72
CHST10	11.80	1.72	2.09
CHST11	10.60	2.04	1.77
CHST12	11.18	2.16	1.08
CHST13	14.64	7.88	3.38
COL18A1	21.08	3.37	2.88
COMT	17.34	2.65	2.99
CROT	8.13	2.46	1.13
CRYZ	12.21	3.20	3.05
CTSK	16.80	2.80	4.60
CYB5R3	11.69	1.85	1.97
CYP1A1	13.14	6.79	0.88
CYP1A2	10.22	3.00	2.29
CYP1B1	8.78	2.34	1.17
CYP2A6	**28.96**	7.00	2.86
CYP2A7	7.46	1.70	0.24
CYP2A13	20.76	6.84	2.74
CYP2B6	17.41	3.21	1.18
CYP2C18	14.32	4.17	1.09

CYP2C19	8.64	1.63	1.09
CYP2C8	10.94	3.04	1.22
CYP2C9	11.75	2.08	1.79
CYP2D6	27.50	**8.80**	2.93
CYP2E1	9.03	1.32	2.10
CYP2F1	16.44	4.62	2.03
CYP2J2	8.24	2.69	0.84
CYP2R1	7.30	3.34	1.98
CYP2S1	13.19	2.97	1.49
CYP3A4	8.45	2.06	1.08
CYP3A5	4.72	0.85	1.23
CYP3A7	6.73	0.94	1.65
CYP3A43	10.55	3.05	0.89
CYP4A11	12.79	1.96	1.20
CYP4B1	14.32	4.44	2.72
CYP4F2	14.18	3.30	0.99
CYP4F3	22.82	3.50	3.22
CYP4F8	20.84	5.60	3.86
CYP4F11	11.89	2.35	0.73
CYP4F12	22.59	4.03	1.48
CYP4Z1	10.88	2.01	0.60
CYP7A1	11.27	3.69	2.40
CYP7B1	10.89	3.40	2.27
CYP8B1	19.75	5.06	3.04
CYP11A1	5.69	1.22	1.36
CYP11B1	17.35	1.99	2.17
CYP11B2	17.49	3.54	1.97
CYP17A1	7.70	0.89	1.07
CYP19A1	6.93	1.49	1.30
CYP20A1	5.92	1.26	1.40
CYP21A2	18.08	2.81	1.25
CYP24A1	13.30	2.55	1.75
CYP26A1	13.70	**9.05**	0.73
CYP26C1	17.88	7.70	3.03
CYP27A1	9.83	4.18	1.46
CYP27B1	7.42	2.33	0.64
CYP39A1	7.01	2.30	0.60
CYP46A1	9.60	0.64	2.24
CYP51A1	10.32	2.68	1.47
DCK	9.32	1.94	1.16
DDO	17.10	5.03	1.68
DHRS1	10.29	2.57	2.73
DHRS2	9.64	1.97	1.28

DHRS3	11.43	2.60	2.08
DHRS4	11.75	3.31	1.66
DHRS4L2	12.87	4.61	2.17
DHRS7	12.02	2.88	2.72
DHRS7B	7.44	1.06	1.46
DHRS7C	16.17	4.04	2.02
DHRS9	9.65	3.13	0.88
DHRS12	9.88	1.34	1.71
DHRS13	8.71	1.88	2.83
DHRSX	7.08	1.57	2.75
DNTTIP2	17.68	2.67	3.00
DPEP1	15.35	2.53	2.72
DPYD	9.44	4.84	2.48
DRD2	12.88	2.43	3.87
DRD3	8.45	2.07	0.59
EAF2	8.18	2.31	1.68
EPHX1	22.16	3.61	5.09
EPHX2	12.92	2.26	2.53
ERCC2	15.85	5.28	1.06
EXOC6	10.12	3.10	1.86
F2	10.92	1.94	0.18
F3	9.93	1.53	1.91
F5	12.13	3.24	1.34
F7	11.93	3.05	1.25
F8	3.86	0.94	0.99
F9	4.61	0.71	1.06
F10	20.09	3.69	2.26
F11	10.81	1.57	1.26
F12	10.88	3.92	0.44
F13A1	10.22	2.51	1.39
F13B	8.91	3.13	1.25
FCER1G	27.85	1.61	**7.50**
FGB	11.71	4.24	3.22
FMO1	10.64	2.86	2.46
FMO2	9.97	2.55	1.22
FMO3	9.72	1.57	1.57
FMO4	8.69	1.84	2.01
FMO5	10.24	1.89	2.02
FMO6P	10.02	2.31	1.54
FTO	12.80	1.82	3.37
G0S2	16.90	2.82	1.88
G6PD	7.83	2.50	1.57
GCLC	9.05	1.42	1.57

GCLM	8.58	1.16	0.99
GP1BA	15.37	5.12	1.46
GP5	13.18	3.62	1.29
GP6	20.07	2.95	2.16
GP9	12.76	3.00	0.75
GPLD1	16.59	1.74	2.01
GPX1	18.79	**10.25**	3.42
GPX2	10.62	4.08	2.04
GPX3	12.02	1.63	1.43
GPX4	12.48	4.41	0.73
GPX5	9.64	1.75	1.40
GPX6	17.49	2.96	3.26
GPX7	11.54	4.12	2.06
GRIN2B	9.92	4.40	0.94
GSR	7.51	1.52	0.94
GSS	5.32	2.13	0.76
GSTA1	13.75	1.76	2.93
GSTA2	16.20	4.26	2.27
GSTA3	15.65	4.27	3.56
GSTA4	11.12	1.59	1.59
GSTA5	13.07	3.53	0.71
GSTCD	6.63	0.98	1.64
GSTK1	7.50	1.67	1.67
GSTM1	27.05	1.79	3.32
GSTM2	13.77	1.40	1.90
GSTM3	9.43	1.10	2.04
GSTM4	12.92	1.78	2.23
GSTM5	17.46	2.52	2.85
GSTO1	13.28	2.72	0.91
GSTO2	9.18	1.22	1.35
GSTP1	13.66	1.66	0.41
GSTT1	6.56	1.46	2.55
GSTT2	4.06	NA	NA
GSTZ1	9.89	1.34	1.75
HAGH	20.86	2.52	2.28
HLA-DOB	19.29	2.89	0.96
HMGCR	5.43	1.26	1.26
HNF4A	8.59	2.38	0.93
HNMT	7.26	1.70	1.70
HSD11B1	7.66	1.92	1.09
HSD17B11	8.75	2.27	1.46
HSD17B14	18.66	4.94	**5.21**
HTR2A	8.24	2.12	0.59

IAPP	9.69	2.04	2.80
IGF2R	10.45	2.71	1.69
IL10	10.45	1.98	2.54
INTS12	8.72	1.68	2.91
ITGA2	9.63	1.80	0.98
ITGB3	8.32	1.78	1.86
KCNH2	12.33	4.55	2.71
KCNJ11	14.16	4.80	1.01
KCNJ6	4.75	1.43	0.48
KLKB1	11.32	1.85	1.23
LDLR	12.07	2.49	1.24
LPA	8.74	2.37	1.05
LPL	12.50	2.75	0.75
MAOA	3.90	0.53	0.97
MAOB	4.97	1.08	0.32
MAT1A	12.60	2.61	1.88
METAP1	5.76	1.03	1.54
MGST1	9.84	0.43	2.38
MGST2	8.70	1.57	1.85
MGST3	15.33	1.07	3.42
MPO	10.01	3.75	1.88
MTHFR	10.47	2.90	1.30
NAT1	13.23	1.96	2.94
NAT2	16.92	8.15	1.25
NFE2L2	7.07	1.52	1.79
NHLRC1	11.25	2.34	0.94
NNMT	9.17	2.24	1.02
NOS1	9.28	1.71	1.87
NOS3	16.37	5.09	2.33
NQO1	10.21	3.33	2.66
NR1D1	9.87	5.14	1.23
NR1H4	7.80	1.98	1.40
NR1I2	10.65	2.11	1.51
NR1I3	15.31	4.46	1.94
NR3C1	6.14	2.27	2.09
ORM1	16.66	0.55	2.19
ORM2	**30.18**	1.40	4.21
P2RY1	7.40	2.90	1.29
P2RY2	13.15	4.23	3.76
P2RY12	8.79	1.71	0.98
PDE3A	12.36	2.86	2.57
PDE3B	4.94	1.54	1.12
PKD2	6.64	1.88	0.90

PLG	10.92	1.72	2.39
PNMT	13.56	7.05	2.17
PON1	13.01	2.60	1.99
PON2	8.04	2.28	1.20
PON3	11.08	1.99	0.71
POR	13.68	3.86	2.35
PPARA	9.08	0.89	1.61
PPARD	9.70	2.75	2.39
PPARG	7.38	1.30	1.95
PPP1R9A	8.86	2.61	2.29
PRKAB2	9.09	1.61	1.04
PROS1	6.78	1.47	1.57
PROZ	11.52	1.51	0.94
PRSS53	8.04	3.17	1.59
PSMB8	**38.92**	7.33	2.54
PTGIS	10.50	2.26	1.69
PTGS2	8.78	1.98	1.84
RALBP1	7.65	1.72	2.20
RPL13	19.17	4.06	3.74
RXRA	8.03	0.84	1.57
SCN5A	11.92	4.90	1.39
SCUBE1	12.34	3.19	1.96
SELP	13.48	2.25	2.15
SERPINA7	3.88	1.55	0.78
SERPINA10	9.58	1.97	1.23
SETD4	10.83	1.83	2.37
SGK1	9.94	2.37	1.93
SGOL2	9.28	1.98	1.15
SHBG	9.03	1.19	0.48
SLC2A4	11.41	5.53	1.38
SLC2A5	9.73	1.36	2.33
SLC5A6	7.94	1.74	1.41
SLC6A6	8.68	1.56	1.48
SLC7A5	13.32	5.24	3.19
SLC7A7	14.05	2.84	1.68
SLC7A8	8.31	2.28	2.01
SLC9B2	8.79	2.45	1.94
SLC10A1	14.34	5.92	1.25
SLC10A2	12.56	3.52	0.50
SLC13A1	9.04	1.10	1.74
SLC13A2	11.00	3.72	2.26
SLC13A3	7.54	1.40	0.93
SLC15A1	8.95	1.92	0.55

SLC15A2	7.80	1.66	1.04
SLC16A1	8.59	2.42	2.66
SLC19A1	21.09	3.85	3.79
SLC22A1	15.57	5.65	1.88
SLC22A2	12.08	3.24	1.90
SLC22A3	9.74	2.47	1.10
SLC22A4	8.40	2.44	1.35
SLC22A5	10.26	2.30	1.80
SLC22A6	12.46	4.57	1.45
SLC22A7	10.40	2.77	1.39
SLC22A8	6.34	1.82	1.15
SLC22A9	14.41	2.43	1.87
SLC22A10	14.19	3.24	1.68
SLC22A11	12.17	2.61	1.59
SLC22A12	10.83	1.88	0.94
SLC22A13	10.69	5.00	0.45
SLC22A14	18.59	2.72	1.51
SLC22A15	6.50	1.68	0.69
SLC22A16	15.19	3.31	0.55
SLC22A17	9.49	5.18	1.38
SLC22A18	14.49	1.45	2.28
SLC22A18AS	16.24	2.15	3.34
SLC27A1	10.36	2.44	1.07
SLC28A1	13.41	2.18	1.22
SLC28A2	8.06	2.07	2.39
SLC28A3	11.48	2.42	2.26
SLC29A1	7.11	1.42	0.90
SLC29A2	9.08	2.81	1.16
SLCO1A2	9.73	1.81	1.99
SLCO1B1	12.44	2.33	2.24
SLCO1B3	12.47	2.45	0.98
SLCO1C1	10.94	1.96	1.22
SLCO2A1	10.63	2.07	1.69
SLCO2B1	8.01	1.06	1.06
SLCO3A1	10.39	1.88	1.88
SLCO4A1	17.71	4.30	1.69
SLCO4C1	9.38	1.56	1.56
SLCO5A1	10.34	3.41	2.56
SLCO6A1	13.80	1.30	1.69
SOD1	10.34	3.71	1.86
SOD2	15.81	2.33	3.63
SOD3	17.60	4.14	4.66
SPG7	13.72	3.14	1.98

SRD5A2	16.16	3.13	4.17
STK19	10.87	2.17	1.52
SULF1	7.49	2.23	1.44
SULT1A1	17.50	2.25	1.99
SULT1A2	17.12	2.85	1.07
SULT1A3	9.10	NA	3.03
SULT1B1	12.50	3.06	1.53
SULT1C2	7.89	1.56	1.65
SULT1C4	12.22	1.82	2.86
SULT1E1	8.75	0.85	1.49
SULT2A1	10.27	1.28	0.77
SULT2B1	15.80	4.07	0.96
SULT4A1	9.65	1.23	1.93
SV2B	8.88	1.63	1.69
TAP1	23.28	4.04	1.39
TAP2	24.35	3.98	2.35
TBXAS1	9.69	1.80	2.34
TMEM63A	11.88	2.02	2.23
TNF	11.17	1.72	1.72
TNIP1	11.30	1.99	1.55
TOMM40L	8.05	2.50	0.98
TPMT	8.71	1.14	2.86
TPSG1	24.81	4.71	3.45
TSPO	13.43	4.70	2.69
TTBK1	7.66	3.75	1.13
TTR	9.82	2.15	1.84
TYMS	14.26	2.34	1.70
UGT1A1	14.36	2.34	2.84
UGT1A3	11.32	2.38	1.99
UGT1A4	13.94	3.14	2.09
UGT1A5	13.60	3.38	1.76
UGT1A6	13.88	3.11	1.56
UGT1A7	12.78	2.85	1.35
UGT1A8	12.87	2.90	1.35
UGT1A9	12.77	2.75	1.40
UGT1A10	12.65	2.85	1.38
UGT2A1	14.95	5.26	3.88
UGT2B4	13.33	2.18	1.09
UGT2B7	16.64	3.84	0.77
UGT2B10	10.51	3.69	0.99
UGT2B11	25.14	8.38	4.85
UGT2B15	13.74	3.50	1.62
UGT2B17	14.39	3.82	2.06

UGT2B28	12.53	3.68	0.37
UGT8	13.49	5.25	2.25
UNC93B1	10.07	2.20	1.65
UROC1	13.44	2.76	1.14
VDR	10.77	0.91	3.38
VKORC1	24.17	**8.57**	4.90
XDH	12.08	3.39	0.77
ZBED1	7.55	2.27	3.78