Universidad de Oviedo

PROGRAMA DOCTORADO CIENCIAS DE LA SALUD

UNIVERSIDAD DE OVIEDO

DIRECCION DOCTORES: D. VICENTE MARTíN SÁNCHEZ

D. FERNANDO SÁNCHEZ LASHERAS

TUTOR: DOCTOR D. MARCELINO CUESTA IZQUIERDO

OBESIDAD. ACTIVIDAD FÍSICA Y ADHERENCIA A LA DIETA MEDITERRÁNEA EN EL CÁNCER DE PRÓSTATA. INTERACCIÓN DE LA ENFERMEDAD CON EL MEDIO AMBIENTE Y VÍAS GENÉTICAS

AUTOR: D. DAVID ÁLVAREZ GUTIÉRREZ

Universidad de Oviedo

RESUMEN DEL CONTENIDO DE TESIS DOCTORAL

1.- Título de la Tesis	
Español/Otro Idioma: OBESIDAD. ACTIVIDAD FÍSICA Y ADHERENCIA A LA DIETA MEDITERRÁNEA EN EL CÁNCER DE PRÓSTATA. INTERACCIÓN DE LA ENFERMEDAD CON EL MEDIO AMBIENTE Y VÍAS GENÉTICAS	Inglés: OBESITY. PHYSICAL ACTIVITY AND ADHERENCE TO THE MEDITERRANEAN DIET IN PROSTATE CANCER. INTERACTION OF THE DISEASE WITH THE ENVIRONMENT AND GENETIC PATHWAYS

2.- Autor	
Nombre: David Álvarez Gutiérrez	
Programa de Doctorado: Ciencias de la Salud	
Órgano responsable: Centro Internacional de Posgrado	

RESUMEN (en español)

En la actualidad, se sabe que existen variantes genéticas que pueden emplearse como predictores de la incidencia y pronóstico del cáncer de próstata. El uso para estos propósitos de los polimorfismos de un solo nucleótido (SNP) es una de las áreas de investigación más prometedoras en la investigación del cáncer. El objetivo de este proyecto de investigación es el desarrollo de metodologías para estudiar la influencia de las variantes genéticas (SNP) con la ayuda de diferentes algoritmos de aprendizaje automático.

En esta investigación, se han desarrollado y probado algunas nuevas metodologías de aprendizaje automático con datos obtenidos de la base de datos de MCC-Spain seleccionando casos y controles como un grupo heterogéneo.

Este trabajo presenta una nueva metodología para GWAS que hace uso de los algoritmos denominados extreme learning machine y differential evolución. La metodología propuesta se probó con la ayuda de la información genética (370.750 polimorfismos de un solo nucleótido) de 2049 individuos, de los cuales 1076 padecían la enfermedad. Se probó la posible relación de 10 vías diferentes con esta enfermedad. Los resultados obtenidos mostraron que la metodología propuesta es adecuada para detectar rutas relevantes para el rasgo bajo análisis con un costo computacional menor que otras metodologías de aprendizaje automático propuestas anteriormente.

RESUMEN (en Inglés)

At present, it is known that there are genetic variants that can be used as predictors of the incidence and prognosis of prostate cancer. The use of single nucleotide polymorphisms (SNPs) for these purposes is one of the most promising research areas in cancer research. The objective of this research project is the development of methodologies to study the influence of genetic variants (SNPs) with the help of different machine learning algorithms.

In this research, some new machine learning methodologies have been developed and tested with data obtained from the MCC-Spain database selecting cases and controls as a heterogeneous group.

This work presents a new methodology for GWAS that makes use of the algorithms called extreme learning machine and differential evolution. The proposed methodology was tested with the help of genetic information (370,750 single nucleotide polymorphisms) of 2049 individuals, of which 1076 had the disease. The possible

Universidad de Oviedo

relationship of 10 different pathways with this disease was tested. The results obtained showed that the proposed methodology is adequate to detect relevant pathways for the trait under analysis with a lower computational cost than other previously proposed machine learning methodologies.

SR. PRESIDENTE DE LA COMISIÓN ACADÉMICA DEL PROGRAMA DE DOCTORADO EN CIENCIAS DE LA SALUD

CONCLUSIONES

- La investigación descrita en este trabajo presenta un algoritmo novedoso basado en metodologías de aprendizaje automático que ha demostrado dar un buen rendimiento en GWAS. Este trabajo continúa una línea de investigación [12] que hace uso de algoritmos y combina diferentes metodologías de aprendizaje automático. Uno de los principales inconvenientes de estas metodologías es la falta de una explicación biológica simple de los resultados obtenidos, ya que, aunque existen muchas vías cuyas relaciones con la enfermedad y los rasgos son bien conocidas, es difícil encontrar cómo se comporta cada uno de los SNP que forman la vía en el proceso e influye en él. A pesar de ello, es posible explicar la influencia de las diferentes vías sobre el cáncer colorrectal haciendo uso de la literatura disponible.

- Como ya se ha dicho en un trabajo anterior [12], en nuestra opinión, debido a la falta actual de una metodología multivariante gold estándar para GWAS, todos los algoritmos, como el presentado en esta investigación, deben tenerse en cuenta en GWAS. Además, en comparación con el algoritmo anterior propuesto por los autores [12], el hecho de que el tiempo de cálculo requerido para este sea de aproximadamente 1 dividido por 47 debe considerarse como una de las principales ventajas del algoritmo propuesto en esta investigación. Este resultado se debe principalmente al rápido entrenamiento del ELM que ya fue declarado en la literatura [62].

- Además, como se demostró en la presente investigación, y en línea con trabajos anteriores, el aprendizaje automático es una herramienta valiosa para el análisis GWAS, ya que es capaz de encontrar SNP y los loci de interés. A pesar de ellos, uno de los principales inconvenientes del machine learning es la falta de una explicación clara de los modelos obtenidos con la mayoría de las metodologías. Este hecho es importante en el campo de GWAS, donde las relaciones entre genes y rasgos son a menudo difíciles de interpretar. Teniendo en cuenta lo mencionado anteriormente, una de nuestras futuras líneas de investigación consistirá en aplicar un algoritmo explicable de aprendizaje automático, como el bosque

aleatorio, al problema en estudio en este trabajo. Adicionalmente, y con el fin de contribuir a consolidar el papel de los algoritmos de machine learning en GWAS, los autores también se centrarán en comparar los resultados obtenidos con otras metodologías que son comunes en GWAS y que no pertenecen al campo del machine learning con el fin de cerrar la brecha entre los diferentes enfoques de GWAS.

ÍNDICE GENERAL

ARTÍCULO I

Sección 1.01. OBESIDAD

La obesidad conceptualmente es un acúmulo energético excesivo en forma de grasa en el panículo adiposo. Se trata de una acumulación en exceso de depósitos de grasa en un organismo, que resulta de un desequilibrio, sostenido en el tiempo, entre la ingesta y el gasto energético. En la década pasada, este sería el enfoque inicial, casi un enfoque matemático entre ingresos y gastos; sin embargo, las perspectivas de la investigación han experimentado un cambio importante en base al desarrollo de las ciencias relacionadas con la biología. La Nutrigenómica supone un cambio de visión que aúna conocimientos en genómica como transcriptómica, proteómica, metabolómica junto con la bioinformática y la biología molecular (A. Palou, 2004).

La Organización Mundial de la Salud (OMS) define como **obesidad** cuando el índice de masa corporal (**IMC**, cociente entre el peso y la estatura de un individuo al cuadrado) es **igual o superior a 30 kg. / m2**.El código de clasificación internacional es: **E66**(CIE 10) y**T82** (CIAP2).

La obesidad ha sido ya reconocida a nivel mundial como un grave problema de salud global, que ha alcanzado las cifras de auténtica epidemia iniciada en el siglo XX y que, de no realizar importantes cambios, será también del siglo XXI. En la actualidad uno de cada 4 españoles mayores de 18 años es obeso (Gutiérrez-Fisac, 2012).

El sobrepeso y la obesidad son el quinto principal factor de riesgo de fallecimiento en el mundo. Cada año fallecen 2.8 millones de personas adultas debido a este problema.La curva de mortalidad en relación al peso es de tipo “J”, correspondiendo la menor al índice entre 21 y 25. La esperanza de vida disminuye en relación inversa al sobrepeso.

En cada individuo en concreto, está relacionado con multitud de desórdenes y enfermedades que paso a enumerar de forma escueta: *síndrome metabólico, diabetes tipo II* (la obesidad es el factor de riesgo más importante para desarrollar una diabetes tipo II), enfermedades cardiovasculares: *hipertensión arterial* (la prevalencia de la HTA en obesos es 10 veces más que en población con normopeso), i*nfarto agudo de miocardio, aterosclerosis precoz, esterilidad*, alteraciones con el sueño (*síndrome de apnea obstructiva del sueño*), alteraciones hepáticas (*esteatosis hepática*), etc. Además, existen una serie de patologías malignas (cánceres) relacionadas con el mismo: en los varones (próstata y colorrectal) y en las mujeres (endometrio, vías biliares, mama y cérvix).

Durante el embarazo mayor incidencia de *HTA, diabetes gestacional, toxemia gravídica y problemas obstétricos.*

Ha sido claramente documentada en la bibliografía internacional la ***asociación inversa entre actividad física y riesgo de cáncer*** (Nuñez, 2018). Existe una evidencia muy importante entre cáncer de colon y actividad física. Esta relación es moderada en mama, pulmón, endometrio y páncreas y finalmente es débil en próstata y ovario.

La actividad física existe evidencia de que protege de alguna manera frente al ulterior desarrollo del cáncer en general y de algunos en particular; sin embargo, sabemos que por diversos mecanismos y algunos desconocidos. Es conocido que el ejercicio físico altera el perfil hormonal de un individuo reduciendo el estrés oxidativo, desarrolla la función del sistema inmunológico, regula los niveles de insulina, cambia por completo el perfil lipídico. Insulina e IGF1 (insulingrown factor tipo 1) se han relacionado en el desarrollo de cáncer de próstata en pacientes obesos.Durante cualquier proceso de tipo quirúrgico, presentan mayor riesgo de presentar un tromboembolismo pulmonar, infecciones y complicaciones con la anestesia.

Sección 1.02 ACTIVIDAD FÍSICA.

La dieta incorrecta como la inactividad son dos factores de riesgo evitables y modificables; conocer la importancia de cada uno de ellos es ahora un punto de interés de la salud pública. En relación al tema que nos ocupa, sólo es posible la intervención en el estilo de vida con dieta y ejercicio, pero no tienen ambas el mismo impacto sobre la misma. Según un reciente metaanálisis, es más importante una dieta mediterránea (hipocalórica con pocos carbohidratos e hiperproteica) que el ejercicio aeróbico de baja intensidad (también llamado quemagrasas“fatburn”). El ejercicio sí es capaz de contribuir, pero llegado un punto, no seguiría haciendo la progresión en la pérdida de peso a diferencia de una dieta estricta hipocalórica e hiperproteica. No obstante, no cabe duda que las intervenciones a este nivel han de ser conjuntas con una implementación clara de la dieta mediterránea y el ejercicio saludable.

No es menos cierto que existe un tipo de ejercicio que podríamos considerar no saludable o pernicioso que sería aquel que no está adaptado a la condición física del individuo, llevando el cuerpo a un estrés elevado fuera del rango aeróbico para convertir este en anaeróbico con los riesgos que ello conlleva. Un ejercicio físico desproporcionado en volumen e intensidad generaría un mayor estrés oxidativo y, por lo tanto, radicales libres también relacionados en la etiopatogenia del cáncer de las diferentes estirpes celulares.

Es decir, ejercicio físico sí, pero en la cantidad y calidad prescritos para la forma física del individuo. Recientes trabajos también sugieren que la actividad física y el IMC se comportarían realmente como factores de riesgo independientes en algunos tipos de cáncer, lo cual apoyaría claramente que la intervención ha de ser conjunta con dieta y ejercicio para obtener así un mayor impacto sobre la salud. Otra de las preguntas clave en este asunto, es cuándo es realmente útil la intervención por grupo de edad. La respuesta es que cuanto antes mejor, es decir, en la infancia-adolescencia en la lucha contra la obesidad infantil y también en el adulto joven, aunque esta intervención es rentable a lo largo de toda nuestra existencia.

Sección 1.03 BENEFICIOS DE LA DIETA MEDITERRÁNEA.

En el cáncer de próstata existen importantes diferencias geográficas en cuanto a su incidencia y mortalidad, que obedecen a múltiples factores como son genéticos-raciales, migraciones, diferentes dietas, etc.

Las dietas occidentales se caracterizan por elevados porcentajes de aporte calórico y proteico generando mayor mortalidad prematura que la dieta mediterránea (DM) y que las orientales. *La dieta mediterránea se caracteriza por* ***el*** **consumo elevado de frutas y verduras,*consumo elevado de ce*reales y legumbres***, uso de aceite de oliva como lípido fundamental, consumo regular de pescado de pequeño tamaño y procedimientos de hervir en agua o freír en baño de aceite de oliva. Además, consumo tradicional de vino en cantidades moderas en las comidas, escaso aporte cárnico predominando las aves de corral e ingesta moderada baja de productos lácteos, escasez de hidratos simples y casi nula de alimentos industriales*. (J. Ferrís-Tortajada, 2012).

La ***Granada (Punicagranatum)*** es una fruta que se comporta como un potente antioxidante, que es capaz de inducir apoptosis e impedir las metástasis de las células tumorales prostáticas (Yuanle Deng et al. 2017).

El ***Brócoli (Brassicaoleracea)*** y la ***Cúrcuma (Curcuma longa)*** también presentan las mismas propiedades antioxidantes y antitumorales (Xingwang Ye, Shilpa N et al. 2012)

Existe una evidencia a favor de las ***dietas elevadas en calcio, la leche y los derivados y las carnes procesadas, roja, ahumada, sazonada,*** etc. aumentan el riesgo de cáncer de próstata.

Las ***legumbres, el selenio, los licopenos, el alfa tocoferol*** se ha demostrado su efecto protector.

En cuanto al ***vino***, asunto que siempre ha generado controversia, existe una clara evidencia de que los polifenoles inhiben la proliferación tumoral e inducen apoptosis en las células del cáncer de próstata. En el vino tinto existe 5 polifenoles: quercetina, morina, rutina, ácidos gálico y tánico. (Ignacio Romero, 2004).

En cuanto al ***té verde*** hay que puntualizar que sería la bebida más consumida en el mundo si descartamos el agua. El té contiene polifenoles siendo los flavonoides los más importantes. Dentro de estos se encuentran: epicatequina, epicatequina-3-galato, epigalocatequina, y epigalocatequina-3-galato. Una inyección diaria de 1 mg de epigalocatequina en ratones atímicos, producía en un estudio una reducción de tumores prostáticos entre 20 y 30 %.

No se ha encontrado una gran asociación entre el ***consumo de pescado*** y el riesgo de cáncer de próstata y mortalidad asociada. Sólo un leve incremento en pescado graso asociado a un riesgo aumentado de este tumor y mortalidad sobreañadida por tal alteración. (MaleneOutzen*et al.* 2018)

Sección 1.04 INTERACCIÓN ENFERMEDAD Y MEDIOAMBIENTE.

Las ***alteraciones en el microbioma*** de la próstata están directamente relacionadas con una inflamación crónica de las células prostáticas, que finalmente desemboca en el cáncer. Es conocido que esa alteración en la flora microbiológica, está directamente relacionado con la inflamación crónica en la zona, además de otras zonas anexas que también ejercerían su interacción múltiple como pueden ser la zona del tracto digestivo y también el urinario anexo. (Porter C.M, 2018). Las ***infecciones víricas*** y, más concretamente, los ***papilomavirus*** se han relacionado a lo largo de la historia reciente en la etiopatogenia de esta enfermedad. Sin embargo, los estudios consultados muestran que aún este asunto es muestra de controversia y no está aún claro pues requiere de más estudios (Volgareva, 2015). ***Tricomonas Vaginalis*** no ha demostrado aumentar tal riesgo de PC (Shui *et al*, 2016). La ***actividad sexual*** y el riesgo del ulterior desarrollo de PC ha sido estudio en múltiples estudios. La conclusión de todos los estudios es la misma y es que la eyaculación frecuente es una evidencia que es un factor protector. Múltiples parejas sexuales es probable que protejan del desarrollo del cáncer excluyendo el riesgo de las ETS sobrevenidas de tales actitudes. Los hombres homosexuales tienen un mayor riesgo para este diagnóstico (Koth*et al*., 2015)

Existe una evidencia manifiesta, que se indica en múltiples estudios, que existe una asociación entre la exposición a ***agentes químicos pesticidas*** utilizados preferentemente en la agricultura y el ulterior desarrollo del cáncer de próstata. Revisando este nivel de evidencia, existen 49 estudios publicados entre 1993 y 2015 donde se analiza en población inglesa este efecto devastador sobre la biología prostática. (Silva JF, *et al.)*Los ***niveles de arsénico*** en el agua de bebida fueron analizados en un reciente estudio en Illinois (Estados Unidos) en el que se relacionó la concentración de este elemento químico con la incidencia de cáncer de próstata, concluyendo que existe una correlación lineal entre ambos fácilmente demostrable (Bulka*et al.* 2016)

La ***exposición al asbesto*** ha sido demostrada que tiene relación con un aumento del riesgo de PC. Existe un metaanálisis, que revisa 17 estudios independientes, que concluye que es un agente etiológico implicado en el desarrollo de este tipo de tumores (Rui Peng*et al.* 2019).

La exposición ocupacional a los ***rayos ultravioleta solares*** y el posterior desarrollo de esta tumoración se ha demostrado una escasa o nula correlación, apuntando en algunos estudios incluso una disminución en el riesgo en aquellos que están expuestos por razones puramente ocupacionales (Peters CE, 2016).

La *exposición a la luz artificial por la noche de alta intensidad* aumenta el riesgo de desarrollo posterior de cáncer de pulmón y también prostático. La alteración en el ritmo circadiano se cree está relacionada con la patogenia de esta alteración. Así se demuestra después del estudio MCCSPAIN (García-Sáenz, 2018).

La *exposición al humo del tabaco* se demuestra que tiene una relación no evidente con el desarrollo de PC. No sólo no existe tal asociación para el desarrollo de la enfermedad, sino tampoco está relacionado con el nivel de agresividad (Cosimo De Nunzio et al. Minerva UrolNefrol 2018)

La *ingesta de alcohol* tanto de forma leve como moderada, no está demostrada su asociación con el ulterior desarrollo de PC (RA Breslow *et al. Nutr*Cancer 1998).

Sección 1.05 BIOGÉNESIS MITOCONDRIAL Y CÁNCER DE PRÓSTATA

Las mitocondrias son organelas citoplasmáticas implicadas en el metabolismo energético, principalmente en la fosforilación oxidativa. La cadena respiratoria mitocondrial está compuesta por cinco complejos y dos moléculas que actúan a modo de nexo de unión o lanzadera, la coenzima Q y el citocromo C. La función mitocondrial está regulada por un doble sistema genético, uno propio, el ADN mitocondrial integrado por 16569 pares de bases que codifica 22 ARN de transferencia, dos ARN ribosómicos y 13 péptidos de la cadena respiratoria. El otro, el ADN nuclear. El ADN mitocondrial se hereda integra y directamente de la madre por lo que tiene un tipo herencia peculiar.

Hace 70 años, "Otto Warburg" observó que las células cancerígenas producían un exceso de lactato en presencia de oxígeno. Los tejidos tumorales metabolizan 10 veces más glucosa a lactato que los tejidos normales. La hipótesis de Warburg sostiene que lo que conduce a la carcinogénesis es una respiración celular defectuosa, causada por un daño mitocondrial. Las células cancerosas fermentan y siguen mutando. Esto les aporta ventajas en la supervivencia. El ácido láctico les confiere propiedades invasivas.

A nivel metabólico, la célula tumoral utiliza una gran cantidad de combustibles como glutamina, glucosa, ácidos grasos, cuerpos cetónicos, acetato…tienen una gran demanda de ácidos grasos y glucosa. En las células cancerosas, el metabolismo es reprogramado para obtener la máxima cantidad de ATP. Ciertos quimioterápicos pueden inducir un daño mitocondrial como son Mitomicina C, Adriamicina (Doxorrubicina), etc.

La relación entre el DNA Mitocondrial alterado y el cáncer está más que demostrada. La eliminación de este ha disminuido la tasa de crecimiento y la formación de nuevos tumores. Las mutaciones en el DNA mitocondrial aumentan el riesgo de cáncer de próstata (sdh, fh, idh1, idh2…)-

Estas características se han utilizado en el diagnósticoTomografía de Emisión de Positrones(PER)con glucosa no metabolizable (18-F-2 desoxiglucosa) que aumenta su captación con tal diagnóstico. Se ha demostrado que la actividad bioenergética correlaciona de forma inversa con la tasa de glucólisis aerobia en carcinoma colorrectal. Por tanto, la actividad mitocondrial en la célula actúa como supresor tumoral. También se ha utilizado en el pronóstico, pues ha demostrado una relación con el grado de actividad metabólica.

Sin embargo, alteraciones en la función mitocondrial pueden provocar cambios en el DNA nuclear y viceversa, a través de pathways y finalmente, estos cambios pueden provocan cambios en el estroma y su metabolismo, constituyendo lo que se ha dado en llamar *"Efecto Warburg Inverso".* Destrucción mitocondrial en los fibroblastos del estroma (mitofagia).

El cáncer recicla nutrientes derivados del catabolismo del estroma que se utilizan para alimentar el tumor, proceso de autofagia o "selfcanibalismo". Este fenómeno explica la caquexia asociada al cáncer (atrofia sistémica) en el que los pacientes son incapaces de mantener su peso corporal. Los nutrientes reciclados producidos por autofagia en las células estromales, proporcionan un flujo constante de metabolitos ricos en energía induciendo biogénesis mitocondrial y la protección de las células cancerosas contra la apoptosis. Se genera un microambiente tumoral letal.

Las aplicaciones clínicas del efecto Warburg inverso serían por un lado como marcador pronóstico con el stromal cav-1 que predice la repetición temprana, metástasis ganglionar y resistencia a los fármacos en prácticamente todos los subtipos de cáncer de mama. Pacientes con cáncer de mama TN con ausencia de estroma tienen una tasa de supervivencia menor al 10% después de 5 años, mientras que con alta stromal cav-1 tienen una supervivencia del 75% a los 5 años.

En investigaciones se ha demostrado mediante un sistema de cocultivo, que la pérdida del estroma Cav-1 puede ser efectivamente protegido por el tratamiento de antioxidantes (tales como la N-acetilcisteina, quercetina y metformina), o con autofagia (inhibidores de la lactocloroquina). Estas drogas son medicamentos aprobados por la FDA. Demostrada actividad antitumoral en modelos preclínicos.

En resumen, las principales características del cáncer en su relación con la mitocondria serían: (tomado de Hanahan y Weinberg, 2011).

1. Autosuficiencia de las señales de crecimiento.
2. Angiogénesis.
3. Promoción de la inflamación.
4. Insensibilidad a las señales de crecimiento.
5. Potencial replicativo ilimitado e invasión.
6. Inestabilidad genómica y mutacional.
7. Evasión de la apoptosis.
8. Metástasis de otros tejidos.
9. Desregulación energética.
10. Evasión del sistema inmune.

Finalmente, recordar que un oncogén es un gen que codifica una proteína capaz de transformar células en cultivo o inducir cáncer. En la mitocondria, se desarrollan genes supresores de oncogenes que se comunican con el núcleo; cuando estos genes pierden su capacidad regulatoria e inhibición lo que aumenta la actividad proteica de los oncogenes generando un estímulo para que la célula cancerígena alcance su supervivencia. Ejemplos de genes que participan en estos procesos: p53, myc, etc. y oncogenes c-myc y akt que se consideran relacionados con la proliferación y supervivencia del cáncer.

Sección 1.06 GENÉTICA Y CÁNCER DE PRÓSTATA

El cáncer de próstata tiene un marcado componente genético, de manera que los *antecedentes familiares* tienen un peso relevante como factor de riesgo y se han descrito polimorfismos genéticos que modifican su incidencia y pronóstico (Blows, 2010; Goh, 2012). Este tumor es uno de los cánceres con mayor influencia genética con 15 % con antecedentes familiares positivos.

De la misma manera, el efecto protector del peso adecuado y la actividad física no está bien esclarecido, pero probablemente se relaciona con resistencia a la insulina, balances hormonales, vitamina D, adipocinas, alcohol, y respuestas inflamatoria e inmunitaria (Barnard, 2007; Leitzman, 2011; Na, 2011; Goh, 2012).

Los polimorfismos implicados en la regulación de todos los mecanismos de biosíntesis, transporte, metabolismo, etc. son potenciales factores de susceptibilidad, pero los efectos de estos genes en el riesgo de este tumor puede ser mediado y modificado a través de su implicación en el balance energético, la obesidad, la actividad física y la dieta.

Los estudios mediante GWAS (Genome-Wide AssociationStudies) han permitido identificar variantes genéticas asociadas al cáncer de próstata, así como modificadoras de los efectos de la actividad física en este tumor; sin embargo, por problemas metodológicos, derivados de la necesidad de establecer exigencias muy elevadas de significación estadística, al realizar comparaciones múltiples, muchas variantes genéticas que no pueden alcanzar esa significación han sido poco estudiadas.

Es una de las más importantes herramientas para intentar desenmarañar los alelos que explican variaciones fenotípicas. La variabilidad fenotípica puede deberse a dos causas: mutaciones puntuales (SNPs) y variaciones estructurales (Frazer et al, 2009). Los SNPs constituyen el cambio estadísticamente más prevalente con 62 millones de SNPs estudiados (2014).

En el momento de escribir esta tesis hay más de 1900 estudios publicados que se registraron en el catálogo público de GWAS mantenido en la Oficina de Genómica de Población de NIH (Hindorff et al.) Hay más de 2000 loci que se han relacionado con enfermedad de una manera sólida (Visscher et al., 2012).

Existen SNPs asociados que se enriquecen para loci de rasgos cuantitativos expresados (eQTL) y se han descrito abundantes pistas de pleiotropía (Nicolae et al, 2010). Un estudio en el año 2008 ajustó una distribución exponencial utilizando el tamaño del efecto de los alelos asociados con la altura y predijo que se necesitarían 93.000 SNP causantes para explicar la heredabilidad del rasgo.

Muchos de ellos tienen eminentemente funciones regulatorias. Las variaciones estructurales (StructuralVariation SV) tienen su causa fundamental en la recombinación de los alelos. La recombinación de los alelos es la segunda fuerza evolutiva importante que aumenta la variabilidad genética (Slatkin, 2008).

Este tipo de estudio se realizó por primera vez en el año 1996 por Risch and Merikangas. Está basado en un método estadístico a genoma completo que presenta un mayor rendimiento en variables con efectos genéticos moderados y baja penetrancia a diferencia de los planteamientos genéticos clásicos de variables Mendelianas con efectos genéticos fuertes con una alta penetrancia. Se ha planteado un trabajo colaborativo global llamado “Hapmap Project” (2003).

El enfoque de este tipo de estudios sería semejante a uno epidemiológico de casos y controles. Los casos serían los enfermos o los expuestos a un factor de riesgo y los controles serían los sanos o los no expuestos al factor de riesgo. Se realiza un análisis estadístico basado en una distribución x2 y a partir de ahí se obtienen los resultados.

Este tipo de estudios no son útiles solamente en los estudios en oncología sino en múltiples campos de la medicina como son: infarto agudo de miocardio y eventos cardiovasculares incluída la muerte prematura, deterioro neurológico, obesidad, diabetes Mellitus tipo II, nutrición e IMC, esquizofrenia, etc.

El***pathwayanalysis*****,** en lugar de estudiar la contribución de cada variante genética a la aparición de la enfermedad, analiza si la contribución acumulada de genes con un denominador biológico común es mayor que la esperada por azar (Li, 2011). Esto permite eliminar muchos de los problemas de las comparaciones múltiples de los GWAS, idenficar loci que podrían no ser encontrados **en <u>análisis de SNPs individuales</u>***(SNP: Polimorfismos de un solo nucleótido),* y además aportar un mejor conocimiento sobre los mecanismos subyacentes a la susceptibilidad al cáncer incluyendo la interacción con aspectos ambientas y comportamentales (Wang, 2007).

Existen <u>dos tipos de SNPS</u>: los no funcionales, cuya presencia no implica alteración de ningún tipo y los funcionales cuya presencia implica una serie de consecuencias. Los SNP funcionales pueden ser de varios tipos:

1. rSNP; **SNP Reguladores**: Implicaciones funcionales sobre la expresión génica. Se encuentran en el promotor.
2. srSNP; **SNP Estructurales**: se encuentran en los ARNm primarios y secundarios afectando a su estructura y función de ARN. Alteran la traducción del ARNm, el corte y el empalme, la eficiencia para potenciar o inhibir el corte y empalme y la estabilidad de los ARNm.
3. cSNP; **SNP Codificantes**: se encuentran en los exones (zonas codificantes). Pueden ser sinónimos y no sinónimos. Podrían afectar a la formación y estructura proteica.

Dado que son diversos los mecanismos por los que la actividad física parece influir en la aparición del cáncer de próstata, existen diferentes vías genéticas de interés. Además, se ha implementado un nuevo campo de trabajo que recibe el nombre de **EPIGENÉTICA** que se puede definir *como el estudio de las modificaciones de DNA pero que no son debidas a cambios en la secuencia del DNA y que son heredables, estudio de factores no genéticos que condicionan el futuro desarrollo de un organismo. Una vez que finalizó del Proyecto Genoma 2003, los científicos nos hemos dado cuenta de que el lenguaje de expresión de los genes es mucho más importante de lo que en un principio se esperaba. Múltiples señales químicas regulan la expresión de los genes. Tipos: metilación del ADN,*

modificación de histonas, acetilación de histonas, metilación de histonas, ARN no codificante, etc.

También existe la posibilidad de silenciar genes por Micro ARN, que sería un mecanismo fisiológico y su alteración funcional podría ser la causa de determinados tumores malignos (Garzón et al., 2006). Existen 1200 asociaciones de SNPs vinculados a 200 enfermedades. La evidencia ha demostrado que existen SNPs y loci asociados a un mayor riesgo de desarrollar un CP.

Por ejemplo, el ***locus 17q12*** contiene un gen que codifica una proteína llamada **FACTOR NUCLEAR DE HEPATOCITOS 1B (HNF1B**) O**TCF2 FACTOR DE TRANSCRIPCION 2**. Son dos alelos asociados a un SNP dentro de este gen; uno de los cuales se asocia a un mayor riesgo de CP y a progresión a diabetes tipo MODY y viceversa.

Algunos alelos SNPs no se asocian a genes identificables como es el caso de la **región 8q24** donde hay más de ***7 loci*** diferentes en el que los alelos SNP se correlacionan con un aumento de hasta un 50%.

Este 8q24 fue descubierto por primera vez y descrito en 2006 como una importante región del CP. El protooncogen c-myc se encuentra a 200Kb. Por encima de 8q24 y se ha propuesto que regula los elementos de esta región.

Por un lado, son de interés ***las rutas de actuación de la resistencia a la insulina y del IGF1*****,** cuyo receptor interactúa con la **vía ras/raf/MAPK** y la **vía PI3K/Akt/mTOR** que son conocidas vías del desarrollo tumoral (Gallagher, 2011).

Relacionadas también con la resistencia a la insulina, se encuentran ***adipocinas tales como la leptina, la adiponectina o el HGF (HepaticGrowth Factor)*** cuyas rutas de producción y actuación son posibles vías genéticas de interés (Calle, 2004).

En resumen, desde el año 2006 se han catalogado múltiples SNP como alelos de riesgo para CP y su número va en aumento. Se ha estimado que ***estos SNP en conjunto explicarían el 30% de la herencia del CP***. Cada alelo por separado confiere un aumento del riesgo individual. La detección de estos alelos de riesgo en cada individuo, en conjunto con la anamnesis y el PSA podrían aumentar la especificidad y sensibilidad en cuanto a las pruebas diagnósticas en CP. Además de mejorar el diagnóstico, también nos ayudarían en la búsqueda de dianas terapéuticas específicas para este. No cabe duda, a raíz de todos estos conocimientos nuevos surge un nuevo enfoque de la Medicina tradicional o clásica y es la llamada **MEDICINA PERSONALIZADA** que podemos definir como*: El uso del conocimiento en relación a un individuo para su susceptibilidad, pronóstico de enfermedad o respuesta a distintos tratamientos en relación a la salud.*

Del mismo modo, **la *interacción de la dieta y la actividad física en el metabolismo de los lípidos***, en la síntesis y metabolismo de ***las hormonas sexuales, de la vitamina D y el calcio***, en las vías de la ***inflamación crónica, y el estrés oxidativo***, y su asociación con diversos tipos de cáncer hace relevante incluir los genes implicados en estos procesos biológicos en el análisis de las vías genéticas (Friedenreich, 2012; Mc Tiernan, 2010; Khandekar, 2011).

MUTACIONES GEN BRCA2 Aumenta el riesgo de cáncer de mama y ovario de tipo hereditario. Aumenta también el de cáncer de próstata hereditario. 13q12-13

MUTACIONES GEN BRCA1, ATM,PALB E INHIBIDORES DE PARP.

17q21

MUTACIONES EN FOXA 1 (TRES TIPOS: FAST, FURIOUS Y LOUD)

GEN HPC1. Asociado a tumor cerebral. 1q24-25

GEN PCaP. Herencia autosómica dominante. 1q42-43

GEN HPCX. Herencia ligada al sexo. Xp 27-28

GEN CAPB. Asociado a tumor cerebral. 1p36.

ARTICULO III. MATERIAL Y MÉTODOS

Para el presente Proyecto de Investigación se dispone de datos obtenidos del estudio de la MCC-SPAIN. En el año 2008, el Centro de Investigación Biomédica en Red Epidemiología y Salud Pública (CIBERESP) del Instituto Carlos III (ISCIII) inició el estudio MCC-SPAIN. Se trata de un estudio epidemiológico multicéntrico de base poblacional que incluye tumores con elevada incidencia en España. La captación de casos y controles se llevó a cabo entre los años 2008 y 2012 en 10 Comunidades Autónomas Españolas. El objetivo del presente Proyecto de Investigación consiste en estudiar las relaciones e interacción existentes entre la obesidad, la actividad física, las vías genéticas así como la adherencia a la dieta mediterránea en el cáncer de próstata.

El estudio MCC-SPAIN es un estudio de casos y controles de base poblacional que desde 2008 hasta la fecha ha reclutado 2172 casos nuevos de cáncer colorrectal, 1749 de cáncer de mama, 1115 de cáncer de próstata, 492 de cáncer de estómago y 550 de leucemia linfocítica crónica, junto con 4059 controles emparejados por edad y sexo (emparejamiento por frecuencia).

En el MCC-SPAIN participan nodos de 10 comunidades Autónomas (Asturias, Barcelona, Cantabria, Guipúzcoa, Granada, Huelva, León, Madrid, Murcia, Navarra y Valencia). Por personal especialmente adiestrado, se ha realizado una entrevista dirigida y recogida de muestras biológicas en el momento de reclutar los casos y los controles. La entrevista, de aproximadamente una hora y media de duración, ha recogido información sobre factores socio-demográficos, datos antropométricos, tabaco y humo pasivo, historia laboral/pesticidas, exposiciones ambientales (UV y otras), actividad física, historia residencial con

énfasis en la exposición a contaminantes del agua, uso de medicamentos e historia médica, aspectos relacionados con la pubertad, historia familiar y cribado y consumo de líquidos, etc.

Además, se les entregó a los pacientes un cuestionario validado de frecuencia alimentaria de aproximadamente 200 ítems. Al final del cuestionario, se tomaron medidas antropométricas y se obtuvieron muestras de sangre, orina, pelo y uña.

Se extrajeron 27 ml de sangre periférica de los participantes, que se alicuotaron en sangre total, plasma, fracción celular para la extracción de ADN y suero y se almacenaron a -80ºC. Se recogió saliva de los sujetos de los que no se llevó a cabo la extracción de sangre con el kit de ADN ORAGENE y se almacenó a temperatura ambiente hasta la extracción del ADN. Se recogieron muestras de ADN para el 96% de los pacientes encuestados (76%sangre y 27% saliva, obteniendo ambas muestras para un mismo individuo siempre que fue posible). Muestras de uñas de los pies y cabellos de los participantes (79% Y 84% respectivamente). En un tercio de los pacientes del estudio, también se recogieron muestras de orina 60ml alicuotadas y congeladas a -80ºC. En todos los hospitales muestras de las biopsias tumorales asociadas.

Además, de los casos se obtuvo información procedente de las historias clínicas sobre los síntomas (tipo y fecha de aparición), los métodos diagnósticos, el estadio tumoral y el tipo histológico. Se ha realizado también un genotipado del exoma utilizando el EXOME ARRAY de Illumina que contiene aproximadamente 300000 SNPS complementados con 6000 SNPS elegidos expresamente por el MCC-SPAIN.

Se ha realizado un análisis de los datos pareados entre los casos (enfermos) y los controles (sanos-libres de enfermedad) a su vez

distribuidos en homocigóticos dominantes (AA), homocigóticos recesivos (aa) y heterocigóticos (Aa) resultando así una **Tabla de contingencia** sobre la que se le realiza todo el análisis matemático. Por una parte, aquellos resultados que son mayores de 5 se analizan mediante el **TEST CHI CUADRADO** mientras que aquellos que son 5 o menos se utiliza un **TEST EXACTO DE FISHER.**

Para representar todos los SNP del genoma y su impacto (valor de P) sobre la enfermedad se ha utilizado el **GRAFICO MANHATTAN BLOT.**

TEST DE CHI CUADRADO DE PEARSON.

Se denomina así a cualquier test en el que el estadístico empleado sea una distribución chi-cuadrado si la hipótesis nula es cierta. Esta prueba fue desarrollada por Karl Pearson en 1900. Es una de las más conocidas y utilizadas para analizar variables nominales o cualitativas, es decir, para determinar si existe independencia o no entre las variables.

Es un método para verificar si las frecuencias observadas son compatibles con la independencia de la variable. La hipótesis nula será que las variables son independientes. En cuanto al método, se calculan los valores que saldrían con la independencia absoluta, lo que se denominan “frecuencias esperadas”, comparándolos con las frecuencias reales de la muestra. Ho Hipótesis Nula Variables Independientes. H1 Hipótesis Alternativa H1 Grado de dependencia.

La prueba chi-cuadrado utiliza una aproximación a la distribución chi-cuadrado para evaluar la probabilidad de una discrepancia igual o mayor a la que exista en los datos y frecuencias esperadas en la hipótesis nula. La exactitud de la valoración depende de

que los valores esperados no sean muy pequeños y que el contraste no sea muy elevado. La prueba se basa en la suma de todas las diferencias entre las frecuencias observadas de una variable y las esperadas de las mismas utilizando la distribución teórica. La diferencia entre observados y esperados se resume en un valor que adopta el estadístico de chi-cuadrado que tiene asociado un valor de P por debajo del cual se acepta o rechaza la hipótesis de la independencia de las variables.

PRUEBA EXACTA DE FISHER

Es una prueba de significación estadística para analizar las tablas de contingencia. Lleva el nombre de su inventor Ronald Fisher y es una clase de prueba matemática llamada "exacta" pues el significado de la desviación de la hipótesis nula se puede calcular con exactitud en lugar de basarse en una aproximación que se hace en el límite del tamaño de la muestra hasta el infinito como se hace en otros modelos matemáticos. La mayoría de los usos que se hacen de este tipo de análisis en sobre tablas de contingencia 2x2. Se utiliza una distribución llamada hipergeométrica. Para la valoración de muestras con tamaños muestrales grandes, lo ideal es utilizar la chi cuadrado pero cuando el tamaño muestral disminuye a 5 o por debajo de 10 (si sólo tiene un grado de libertad). Esta prueba también recibe el nombre de la mujer saboreando té.

Un ***diagrama Manhattan*** es un tipo de diagrama de dispersión utilizado normalmente en Matemática Estadística para mostrar datos con un gran número de puntos muchos de los cuales tienen una amplitud diferente de cero y con una distribución de valores de gran amplitud. Este tipo de gráfico se utiliza en estudios de asociación del genoma completo GWAS para mostrar

aquellos SNP más significativos. El nombre de “MANHATTAN” se debe a la semejanza del mismo con el perfil de los rascacielos que se elevan por encima de los edificios convencionales.

En estos gráficos la representación de las variables se realiza de la siguiente forma:

En el eje de las abscisas se muestran las coordenadas genómicas.

En el eje de las ordenadas se muestra el logaritmo negativo del valor P de asociación de cada polimorfismo de nucleótido SNP.

Por tanto, y en conclusión de lo anterior, cada punto de la gráfica coincide con un SNP único. La asociación más fuerte será la que tiene un valor de P más pequeño. Sin embargo, el logaritmo negativo la asociación más fuerte será aquel que tenga un valor mayor. Los distintos colores de cada bloque muestran la extensión de cada cromosoma.

ARTICULO IV. Resultados

Sección 1. VARIABLES CONTINUAS.

	TOTALES		CASOS		CONTROLES		
Descripción Variable	**Media**	**Desviación Típica**	**Media**	**Desviación Típica**	**Media**	**Desviación Típica**	**P Valor Test**
Edad en años cumplidos	66,328	8,086	66,066	7,332	66,523	8,603	0,042
Number of right cylinders biopsed	4,879	1,797	4,879	1,797	NaN	NA	NA
Number of right cylinders affected	1,932	1,975	1,932	1,975	NaN	NA	NA
% ofaffectedcylinder	15,817	23,913	15,817	23,913	NaN	NA	NA
Number of left cylinders biopsed	4,880	1,769	4,880	1,769	NaN	NA	NA
Number of left cylinders affected	1,766	1,943	1,766	1,943	NaN	NA	NA
% ofaffectedcylinder	14,865	22,811	14,865	22,811	NaN	NA	NA
Primary Gleason score	3,250	0,496	3,250	0,496	NaN	NA	NA
Secondary Gleason score	3,471	0,634	3,471	0,634	NaN	NA	NA
Total Gleason Score	6,721	0,895	6,721	0,895	NaN	NA	NA
Tumoursize	920,115	1.203,165	920,115	1.203,165	NaN	NA	NA
Primary Gleason score on prostate specimen	3,221	0,552	3,221	0,552	NaN	NA	NA
Secondary Gleason score on prostate specimen	3,537	0,708	3,537	0,708	NaN	NA	NA
Años totales trabajados según códigos CNO	44,439	14,214	45,521	14,107	43,633	14,246	0,000

TABLA 1*Edad en años cumplidos, número de cilindros biopsiados lado derecho, número de cilindros afectados lado derecho, porcentaje de afectación de los cilindros lado derecho, número de cilindros biopsiados lado izquierdo, numero de cilindros afectados lado izquierdo, porcentaje de afectación de los cilindros lado izquierdo, Gleason Score Primario, Gleason Score Secundario, Gleason Score Total, Tamaño Tumoral, Primary Gleason Score onProstateSpeciment, Secondary Gleason Score onProstateSpeciment, años totales trabajados según códigos CNO.*

En la Tabla 1, debemos comentar la *ausencia de datos histopatológicos en el grupo de los controles* debido a que no se realizaron las biopsias prostáticas ni las tomas de muestra a ese nivel. Lo importante a reseñar es la homogeneidad a la hora de agrupar los controles con edad media y desviación muy similares con años totales trabajados según códigos CNO también muy semejantes. Es llamativo el que no existe una clara predominancia de las tumoraciones hacia ninguno de los dos lados de la economía corporal.

En los primeros resultados una variable estadísticamente significativa es la *"edad en años cumplidos"* con un P VALOR 0.042 con una media en los casos de 66.06 y desviación típica de 7,33 mientras que en los controles sale una media de 66.52 con una desviación típica de 8,60. En resumen, los casos son más jóvenes que los controles de forma significativa.

Es llamativo también el dato de la variable *"años totales trabajados según códigos CNO"* coincidiendo de esta manera la media en los casos de 45,52 con desviación típica de 14,10 mientras que los controles tienen 43,63 con una desviación típica de 14,24. En resumidas cuentas que los años totales trabajados se relacionan con una mayor probabilidad del desarrollo ulterior de un cáncer de próstata.

Descripción Variable	TODOS		CASOS		CONTROLES		
	Media	Desviación Típica	Media	Desviación Típica	Media	Desviación Típica	P Valor Test
Actividad física Ocupacional total autorreferida en el total de la vida	151,386	82,662	160,469	76,542	144,688	86,314	0,000
Altura en cm	169,589	7,050	169,337	6,718	169,776	7,285	0,046
Peso en kg (en el momento de la entrevista)	79,153	11,833	78,787	11,583	79,427	12,012	0,204
Peso en kg (a los 20 años)	66,731	16,400	66,658	11,159	66,789	19,562	0,820
Peso en kg (a los 45 años)	73,813	11,352	73,380	11,527	74,170	11,198	0,077
Peso en kg (1 año antes de dco o entrevista)	79,507	12,337	79,244	12,142	79,719	12,492	0,388
Puntuación de Dieta Mediterránea según Sofi	9,033	2,254	8,950	2,229	9,095	2,272	0,395
Puntuación de Dieta Mediterránea según Trichopoulou	4,256	1,650	4,220	1,641	4,282	1,656	0,545
Puntuación de Dieta Mediterránea según Buckland	8,919	3,401	8,942	3,394	8,016	3,408	0,930
Puntuación de Dieta Mediterránea según Fung	4,000	1,773	3,964	1,761	4,027	1,781	0,476
Puntuación de Dieta Mediterránea según Panagiotakos	34,035	4,411	33,826	4,333	34,189	4,464	0,057
# firstdegreecancer	1,513	0,822	1,587	0,912	1,448	0,728	0,017
# seconddegree cáncer	1,159	0,398	1,463	0,389	1,170	0,407	0,646
# otherdegreecancer	1,490	0,872	1,407	0,780	1,555	0,933	0,034
# firstdegreeprostate	1,143	0,408	1,184	0,453	1,064	0,286	0,013
# seconddegreeprostate	1,048	0,218	1,053	0,229	1,000	0,000	0,871
# otherdegreeprostate	1,091	0,339	1,087	0,354	1,100	0,308	0,659

TABLA 2 *Actividad física Ocupacional Total autorreferida en el total de la vida, altura en cm, peso en kg en el momento de la entrevista, peso en kg a los 20 años, peso en kg a los 45 años, peso en kg un año antes del diagnóstico o entrevista, puntuación de dieta mediterránea según Sofi, puntuación de dieta mediterránea según Trichopoulou, puntuación de dieta mediterránea según Buckland, puntuación de dieta mediterránea según Fung, puntuación de dieta mediterránea según Panagiotakos, firstdegree cáncer, seconddegree cáncer, otherdegree cáncer, firstdegreeprostate, seconddegree cáncer, otherdegreeprostate.*

En la Tabla 2 ciertamente lo que podemos decir a este nivel de los resultados es que el grupo de los casos y el de los controles tienen también unas características grupales muy semejantes con talla y peso muy parejo siendo, eso sí, siempre los casos más ligeros que los controles en todos los grupos de edad (20, 45 y un año antes de la entrevista). En la dieta prácticamente idénticos los grupos.

La variable **"Actividad física ocupacional total autorreferida en el total de la vida"** aparece una media en los casos de 160,46 con una desviación típica de 76,54 mientras que en los controles aparece una media de 144,68 con una desviación típica de 86,31. Este dato es particularmente llamativo pues inicialmente podíamos pensar que la actividad física ocupacional podía ser un factor protector cuando, en realidad, es todo lo contrario siendo un factor asociado al ulterior desarrollo de la enfermedad.

Otro dato interesante y difícil de interpretar es la variable **"altura en centímetros"** diferente altura de los casos respecto a los controles siendo significativamente más bajos los casos que los controles. La media de altura de los casos es 169,337 y con la desviación típica de 6,71 mientras que los controles tienen 169,776 con una desviación típica de 7,28.

Hay una variable que es el **peso corporal** en el diagnostico, un año antes del diagnóstico, a los 20 años y a los 45 años. En otros tumores, este es un dato claramente relacionado con la probabilidad de desarrollo del mismo, predictor de evolución y de recaída.

Sin embargo, en nuestro caso es evidente que ninguna de estas 4 variables es estadísticamente significativo puesto que el P valor en todos los casos supera el 0,05 que tenemos como referente de significación. Los P valor de todas estas variables son respectivamente 0,38; 0,204; 0,077; 0,820.

Otro dato curioso y paradójico "a priori" a la hora de interpretar los resultados de esta tabla es que todos los parámetros relacionados con la **"dieta mediterránea"** aparecen estadísticamente no significativos en unos P valor respectivos que desechan la posibilidad de la relación protectora de esta dieta y el desarrollo de CP: 0,545; 0,930; 0,476; 0,057.

Descripción Variable	TOTALES		CASOS		CONTROLES		
	Media	Desviación Típica	Media	Desviación Típica	Media	Desviación Típica	P Valor Test
PSA-value	11,777	29,077	11,777	29,077	NaN	NA	NA
METS de Actividad Física Ocupacional en el tiempo trabajado según códigos CNO	119,357	60,818	125,814	63,097	114,545	58,623	0,000
METS de Actividad Física Ocupacional en el tiempo trabajado en trabajos con actividad intensa según códigos CNO	23,044	56,709	25,232	59,173	21,412	54,764	0,006
METS de Actividad Física Ocupacional en el tiempo trabajado en trabajos con actividad moderada según códigos CNO	52,510	58,089	57,398	59,673	48,867	56,627	0,000
METS de Actividad Física Ocupacional en el tiempo trabajado en trabajos con actividad leve según códigos CNO	37,579	34,671	36,075	35,295	38,700	34,167	0,049
Actividad física Ocupacional total autorreferida en promedio anual	3,030	1,163	3,051	1,145	3,013	1,177	0,366
Indice de masa corporal 1 año antes de dco o entrevista	27,571	3,800	27,598	3,770	27,549	3,824	0,885
Indice de masa corporal a los 20 años	23,046	3,422	23,178	3,605	22,942	3,269	0,511
Indice de masa corporal a los 45 años	25,596	3,523	25,537	3,577	25,645	3,479	0,309
Perímetro de cadera en cm	104,868	8,186	104,226	7,793	105,297	8,414	0,003
Perímetro de cintura en cm	102,076	10,791	101,922	10,899	102,180	10,720	0,847
Consumo de alcohol en el pasado (g/dia)	30,339	34,683	31,850	36,523	29,209	33,210	0,050
INDICE DE MASA CORPORAL IMPUTADO PARA DATOS FALTANTES A PARTIR DE DATOS DE 1 AÑO ANTES	27,566	3,825	27,589	3,772	27,549	3,864	0,883

TABLA 3 *Valor Psa, Mets de Actividad Física Ocupacional en el tiempo trabajado según códigos CNO, Mets de Actividad Física Ocupacional en el tiempo trabajado en actividad intensa según códigos CNO, Mets de Actividad Física Ocupacional en el tiempo trabajado en actividad moderada según códigos CNO, Mets de Actividad Física Ocupacional en el tiempo trabajado en actividad leve según códigos CON, Actividad física Ocupacional Total autorreferida en promedio anual, índice de masa corporal 1 año antes de diagnóstico o entrevista, índice de masa corporal a los 20 años, índice de masa corporal a los 45 años, perímetro de cadera en cm, perímetro de cintura en cm, consumo de alcohol en el pasado, índice de masa corporal Imputado para datos faltantes a partir de datos de 1 año antes.*

En la tabla 3, en cuanto a la actividad física hay unos datos llamativos pero esperados que se interpretan a nuestro modo de análisis de la siguiente forma: en los grupos de actividad física modera e intensa en METS son los casos muy superiores a los controles a diferencia de la actividad física leve que son los controles superiores a los casos. La actividad física intensa y moderada se relacionan más con los casos y la leve con los controles. El perímetro de cadera y de cintura son superiores en los controles que en los casos. El IMC es inferior en los casos que en los controles. El consumo de alcohol era mayor en los casos que en los controles. En este estudio tampoco tenemos cifras de PSA de los controles.

Sección 2. VARIABLES CATEGÓRICAS

TABLA 4	DESCRIPCION DE LA VARIABLE	Moda		Test independencia Var discretas - Casos/Controles	
Variable V	**Descripción Variable.DV**	**Casos**	**Controles**	**P Valor test**	**Test**
id_study	**Subject ID MCC**				
id_entericos	**Subject ID Entericos**				
entericos	**Entericossubject**	2	2	7,69E-20	Chisq test
hospital	**Hospital**	C	C	1,51E-47	Chisq test
area	**Area**	BCN	BCN	9,41E-13	Chisq test
elegible	**Elegibility**				
d_entrev	**Date of interview**				
a4_d_naix	**Date ofBirth**				
education_basic_final	**EducationLevel - BASIC**	Primary school	Primary school	3,92E-07	Chisq test
newid					
score_SE	**Score de Nivel socioeconómico**	Medio	Medio	1,57E-06	Chisq test
ID_study	**Subject ID MCC**				
casop	**Caso de CP o Control**	Caso_prostata	Control		
fecha_dco	**Date of diagnosis**			NaN	Fisher Test

TABLA 4. Sujetos del estudio, sujetos entéricos Id, sujetos entéricos, hospital, elegible, fecha de la entrevista, fecha de nacimiento, nivel educacional, nivel socioeconómico, sujeto MCC, caso de CP, fecha de diagnóstico.

En la tabla 4 se puede observar que hay varias variables de tipo categórico que son estadísticamente significativas en relación al CP en nuestro estudio. Es evidente que la variable Hospital y área salen claramente significativos al igual que el nivel socioeconómico y también de forma independiente si se ha completado o no la educación general básica.

TABLA 5	DESCRIPCION DE LA VARIABLE	Moda		Test independencia Var discretas - Casos/Controles	
Variable V	**Descripción Variable.DV**	**Casos**	**Controles**	**P Valor test**	**Test**
fumador_presente	**Tabaquismo entrevista**	Exfumador	Exfumador	2,10E-01	Chisq test
fumador_ever	**Tabasquismo vida**	1	1	1,05E-01	Chisq test
fumador	**Tabasquimo 1 año antes**	Exfumador un 1 año antes del dco	Exfumador un 1 año antes del dco	2,00E-01	Chisq test
fumador3	**Tabaquismo 3 años antes**	1	1	5,77E-02	Chisq test
fumador5	**Tabaquismo 5 años antes**	Exfumador desde hace más de 5 años	Exfumador desde hace más de 5 años	7,08E-02	Chisq test
fumador10	**Tabaquismo 10 años antes**	Exfumador >=10 años	Exfumador >=10 años	1,69E-02	Chisq test
tipo_biopsia	**Typeofbiopsy**	1	NaN	NaN	Chisq test
fecha_diagnostico	**Date of diagnosis**		NaN	NaN	Chisq test
tipo_histologico	**Typeofprostatecancer**	1	NaN	NaN	Chisq test
ct	**Tumourstage**	3	NaN	NaN	Chisq test
cn	**Nodesstage**	0	NaN	NaN	Chisq test
cm	**Metastasisstage**	0	NaN	NaN	Chisq test
amico	**D' Amicoclassificationrisk**	2	NaN	NaN	Chisq test
cirug_a	**Surgery as first option for treatment**	1	NaN	NaN	Chisq test
pt	**Pathologicaltumourstage**	6	NaN	NaN	Chisq test
pn	**Pathologicalnodesstage**	0 if 9 = Unknown	NaN	NaN	Chisq test
margenqx	**Surgicalmargins**	1	NaN	NaN	Chisq test
radioterapia	**Radiotherapy**	2	NaN	NaN	Chisq test
r_electiva	**Radiotherapy as first option for treatment**	2	NaN	NaN	Chisq test
radyuvante	**Adjuvantradiotherapy**	2	NaN	NaN	Chisq test
ht	**Hormonotherapy**	2	NaN	NaN	Chisq test

TABLA 5.Tabaquismo entrevista, tabaquismo vida, tabaquismo 1 año, tabaquismo 3 años, tabaquismo 5 años, tabaquismo 10 años, tipo de biopsia, fecha del diagnóstico, tipo de cáncer de próstata, estadiaje tumoral, número de ganglio afectados, nivel metastásico, D'Amico clasificación de riesgo, cirugía como primera opción de tratamiento, estadiaje anatomopatológico, revisión ganglios linfáticos con Anatomía Patológica, márgenes quirúrgicos, Radioterapia, Radioterapia como primera opción de tratamiento, Radioterapia coadyuvante, Hormonoterapia.

En la tabla 5 hay que decir que llama la atención un dato en relación al tabaquismo y es el siguiente. Aún a pesar de que se han valorado de forma independiente el hábito al año, a los 3 años, a los 5 años y a los 10 años, el tabaquismo vida y el tabaquismo en entrevista completa. Lo cierto es que estadísticamente significativo sólo nos sale el tabaquismo más allá de los 10 años. Dicho de otra manera, sólo importa si es un exfumador más de 10 años o no pues el resto de variables no presentan impacto sobre el ulterior desarrollo de la enfermedad.

En cuanto al hábito tabáquico y su relación con del desarrollo posterior de otros tumores como el cáncer de pulmón, presenta un periodo ventana de más o menos 20 años en las distintas casuísticas del cáncer de pulmón. Sin embargo, en relación a nuestro estudio a partir de los 10 años ya aparece como factor carcinogenético de primera magnitud.

TABLA 6	DESCRIPCION DE LA VARIABLE	Moda		Test independencia Var discretas - Casos/Controles	
Variable V	Descripción Variable.DV	Casos	Controles	P Valor test	Test
htneoady	Neoadjuvanthormonotherapy	2	NaN	NaN	Chisq test
htadyuv	Adjuvanthormonotherapy	2	NaN	NaN	Chisq test
htpaliat	Hormonotherapy as first option for treatment (palliative)	2	NaN	NaN	Chisq test
vigilancia	Active surveillance/watchfulwaiting	2	NaN	NaN	Chisq test
quimioterapia	Chemotherapy	2	NaN	NaN	Chisq test
gct	Estadio clinico TNM agrupado	estadio menor o igual a t2a	NaN	NaN	Chisq test
gpsa	PSA categorizada	4-10	NaN	NaN	Chisq test
porcT_AFmedia_altaREC	Codif. Porcentaje de tiempo trabajado con actividad moderada o intensa (>2.5 mets) según códigos CNO	50% o más	50% o más	8,02E-03	Chisq test
porcT_AFbajaREC	Codif. Porcentaje de tiemnpo trabajado con actividad baja (<=2.5 mets) según códigos CNO	50% o más	50% o más	8,88E-03	Chisq test
porcAFmuybaja_bajREC	Recod. % tiempo de trabajo con actividad autorreferida muy baja o baja (1-2)	0	0	3,32E-01	Chisq test
porcAFmuyalta_alta_mREC	Recod. % tiempo de trabajo con actividad autorreferida media o más intensa (3-5)	50% o más	50% o más	3,57E-01	Chisq test
ACTDOMREC	Actívidad física Doméstica Realizada	0	0	1,08E-02	Chisq test
TSTA_REC2	Recod.del tiempo sentado en el momento de la entrevista	2	2	3,72E-01	Chisq test
TST30_REC2	Recod.del tiempo sentado entre los 30 y 39 años	1	1	2,09E-01	Chisq test
TST50_REC2	Recod.del tiempo sentado entre los 50 y 59 años	1	1	3,21E-01	Chisq test
METS10Aâ€”OS_2TOTALREC	Recodificación según los METS de Actividad Física Recreacional	0	0	7,97E-01	Chisq test
METS10Aâ€”OS_2ANDARTOT	Recodificación según los METS de ACTIVIDADES LEVES O CAMINAR	0	0	3,91E-01	Chisq test
METS10Aâ€”OS_2MEDIOTOT	Recodificación según los METS de ACTIVIDADES DE INTENSIDAD MEDIA	0	0	3,64E-01	Chisq test
METS10Aâ€”OS_2ALTOTOT	Recodificación según los METS de ACTIVIDADES DE INTENSIDAD ALTA	0	0	9,06E-01	Chisq test

Tabla 6. Variables: Hormonoterapia Neoadjuvante, hormonoterapia adjuvante, hormoterapia como primera opción, Vigilancia, Quimioterapia, Estadío Clínico TNM (estadio igual o menor a T2a), Psa categorizada. Porcentaje del tiempo trabajado en actividad moderada o intensa (más de 2.5 METS); Porcentaje del tiempo trabajado en actividad baja (igual o menor a 2.5 METS). Porcentaje de actividad autorreferida baja. Porcentaje de actividad autorreferida media o intensa. Actividad física doméstica. Tiempo sentado en el momento de la entrevista. Tiempo sentado entre los 30 y 39 años. Tiempo sentado entre los 50 y 59 años. METS de actividad física recreacional. METS de actividad leve o caminar, media o intensa.

En la tabla 6 es muy llamativo el observar como la *actividad física recreacional, sea cual sea el nivel cualitativo,* (leve, moderada o de alta intensidad) no correlaciona con el posterior desarrollo de la enfermedad.

Sin embargo, el tiempo de *actividad física laboral* tanto de actividad moderada o leve como de actividad intensa sí se relaciona de forma clara con el desarrollo de la misma.

También es muy importante significar que el *tiempo sentado entre los 30 y los 50; entre los 50 y los 60 años* no influye necesariamente en desarrollar CP o no.

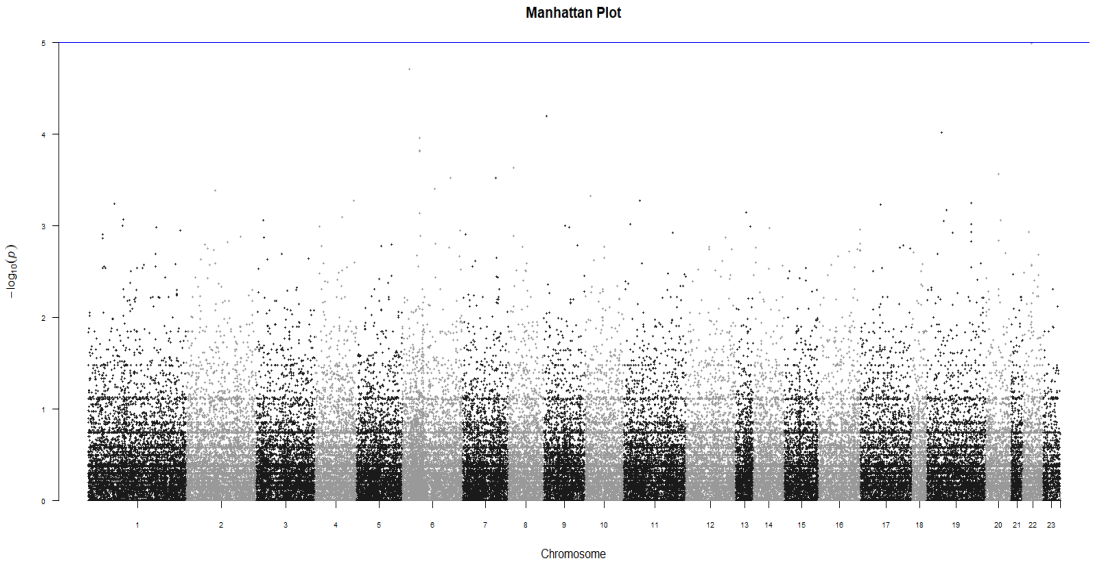
TABLA 7	DESCRIPCION DE LA VARIABLE	MODA			
Variable V	**Descripción Variable.DV**	**Casos**	**Controles**	**P Valor test**	**Test**
IMCh1_rec	**Recodificación de Indice de masa corporal 1 año antes de dco o entrevista**	2 - 25/30	2 - 25/30	4,91E-01	Fisher Test
IMC45rec	**Recodificación deIndice de masa corporal a los 20 años**	1 - 18.5/25	1 - 18.5/25	2,45E-01	Chisq test
IMC20rec	**Recodificacion de Indice de masa corporal a los 45 años**	1 - 18.5/25	1 - 18.5/25	2,36E-01	Chisq test
AINES	**Uso de antiinflamatorios no esteroideos**	No	No	1,11E-03	Chisq test
OHitaly2	**Consumo de alcohol (categorias)**	2	2	5,85E-01	Chisq test
NIVEL_SE	**Score de Nivel socioeconómico**	1	1	1,57E-06	Chisq test
adherenceSofi		0	0	6,42E-01	Chisq test
adherenciaTRICHOPprostata		1	1	8,70E-01	Chisq test
adherenceBUCKLANDprostata		1	1	9,11E-01	Chisq test
adherenceFUNGprostata		1	0	6,07E-01	Chisq test
adherencePANprostata	**Cumplimiento de Dieta Mediterránea según Panagiotakos**	2	1	7,42E-02	Chisq test
FH_cancer	**Familyhistorycancer**	NaN	NaN	NaN	Chisq test
FH_cancer_Prostate	**Familial history of prostate cancer**	No FHca	No FHca	1,40E-22	Chisq test
Fam1_cancer	**Tiene Familiares de 1er grado con cancer (DICOT)**	1	0	1,03E-06	Chisq test
Fam1_CancerProstata	**Tiene Familiares de 1er grado con cancer de próstata (DICOT)**	0	0	5,40E-17	Chisq test

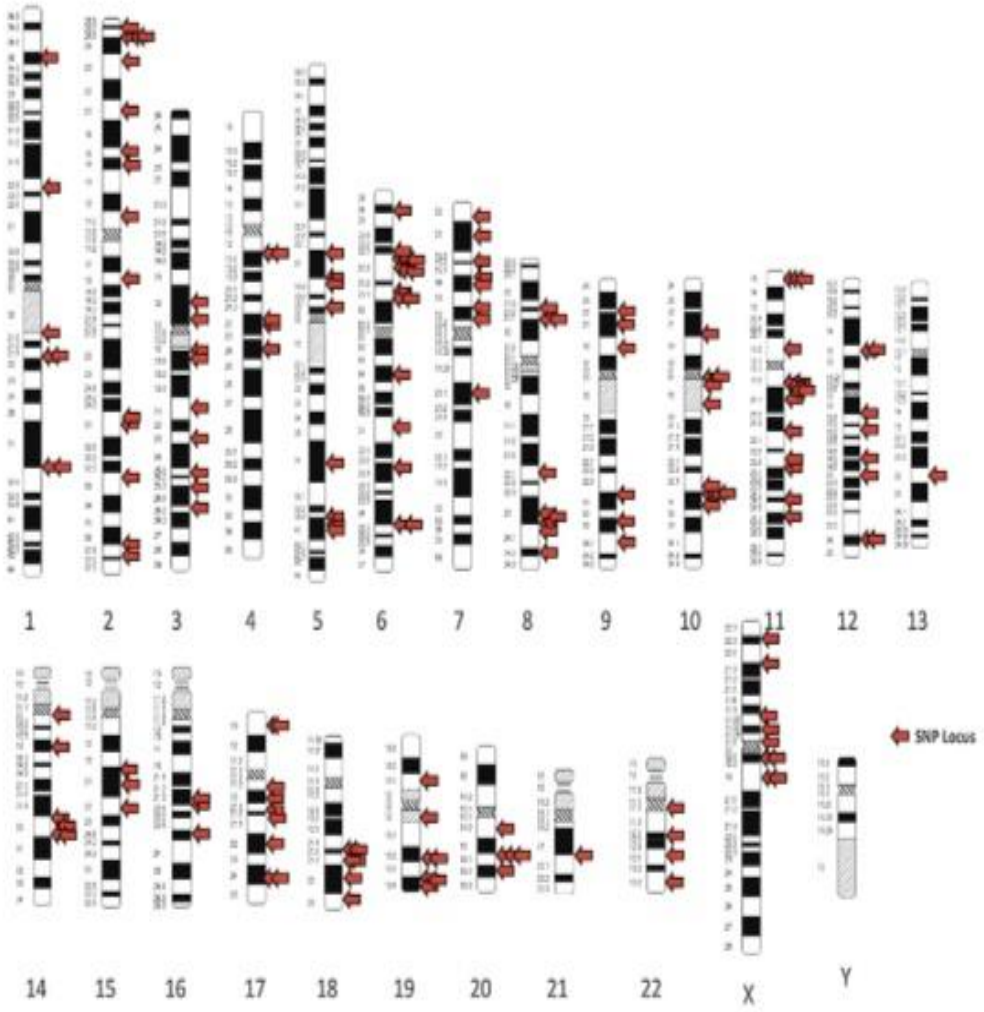
TABLA 7. VARIABLES CATEGÓRICAS. TABLA FINAL

IMC 1 año antes del diagnóstico, IMC a los 20 años e IMC a los 45 años. Uso de AINES. Consumo de alcohol por categorías. Score del nivel socioeconómico. Historia familiar de cáncer; cumplimiento de la dieta mediterránea, historia familiar de cáncer de próstata, familiares de primer grado positivos para cáncer; familiares de primer grado para cáncer de próstata.

En la tabla 7 hemos de decir que de todas aquellas variables antes mencionadas sólo hay 4 en esta tabla que muy significativamente presentan relación directa con el desarrollo del CP. Por un lado, es el consumo de AINES existiendo varias teorías ninguna de ellas confirmada en los estudios ulteriores. En segundo lugar, el nivel socioeconómico y finalmente dos más que, en realidad, podrían interpretarse como una sola, que son la historia familiar de cáncer y la historia familiar de cáncer de próstata. Estos dos últimos, no hacen más que confirmar matemáticamente lo que clínicamente ya intuíamos hace ya mucho tiempo y es, sin duda, la agregación familiar y el componente genético de esta tumoración.

Sección 3. Gráfico Manhattan Plot




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
SNP Locus

SECCIÓN IV:

RESULTADOS ANÁLISIS GENÉTICO:

SNP DE RIESGO EN CÁNCER DE PRÓSTATA

SNP CROMOSOMA 1

rs42457739

Locus 1q32 Alelo de riesgo G Alelo de referencia A

Genes relacionados MDM4, PIK3, C2B.

La proteína MDM4 juega un papel importante en la regulación negativa del supresor tumoral p53 a través de su interacción con MDM2. En línea con esto, se ha observado amplificación de MDM4 en varias formas tumorales. Un polimorfismo (rs4245739 A>C; SNP34091) en la región no traducida DE MDM4 3' se ha informado para crear un sitio de destino para has-miR-191, lo que resulta en una disminución de los niveles de ARNm MDM4. Hay una posible asociación entre MDM4 SNP34091, solo y en combinación con el MDM2 SNP309T>G (rs2279744), y el riesgo de cáncer de mama, colon, pulmón y próstata. Concluimos, estado MDM4 SNP34091 puede ser asociado con un menor riesgo de cáncer de mama, en particular en individuos portadores del genotipo MDM2 SNP309GG, pero no debe estar asociado con cáncer de pulmón, colon o próstata.

SNP CROMOSOMA 2

rs1465618

Locus 2p21 Alelo de riesgo A Alelo de referencia G

Genes relacionados THADA.

Este snp ha sido relacionado con el riesgo de CP aumentado en poblaciones europeas pero no ha sido así en estudios llevados a cabo en población china o en otras poblaciones a nivel mundial. En población europea aumenta la susceptibilidad al CP y también se relaciona en una mayor progresión de dicha patología.

rs13385191

LOCUS 2P24 Alelo de riesgo G Alelo de referencia A

Genes relacionados C2orf43

Cinco polimorfismo de un solo nucleótido (SNP): rs13385191 (cromosoma 2p24), rs12653946 (5p15), rs1983891 (6p21), rs339331 (6p22) y rs9600079 (13q22), en una serie de estudios analizados de control de casos en el Consorcio de Cohortes de Cáncer de Mama y Próstata del Instituto Nacional del Cáncer de Cáncer (BPC3). Se analiza cada SNP para la asociación con el riesgo de cáncer de próstata y se evalúa si las asociaciones difieren con respecto a la gravedad de la enfermedad y la edad de inicio. *Cuatro SNP (rs13385191, rs12653946, rs1983891 y rs339331) se asocian significativamente con el riesgo de cáncer de próstata* (valores P que oscilaban entre 0,01 y 1,1 × 10(-5)). Las frecuencias de alelo y los quirófanos fueron en general más bajos en nuestra población de ascendencia europea que en el descubrimiento de la población asiática. SNP rs13385191 (C2orf43) sólo se asoció con la enfermedad en estadios bajos (P a 0,009, prueba de solo caso). Ningún otro SNP mostró asociación con la gravedad de la enfermedad o la edad de inicio. Cuatro SNP asociados con el riesgo de cáncer de próstata en una población asiática también están asociados con el riesgo de cáncer de próstata en hombres de ascendencia europea.

Rs12621278

LOCUS 2Q31 ALELO DE REFERENCIA A ALELO DE RIESGO G

GENES RELACIONADOS ITGA6

Cinco SNP muestran asociación independiente con progresión del cáncer de próstata (rs12621278, rs629242, rs9364554, rs4430796 y rs5945572) basándose en el análisis de regresión escalonada. *El SNP más fuerte fue rs12621278 en el locus ITGA6, que se asoció con un riesgo de progresión de 2,4 veces mayor (P a 0,0003).* Al considerar la

suma de alelos de riesgo en estos cinco SNP, cada alelo adicional se asoció con un aumento del 29% en el riesgo de progresión (intervalo de confianza del 95%, 1,12-1-47). Encontramos que cinco de los loci de susceptibilidad al cáncer de próstata recientemente destacados, también influyen en la progresión del cáncer de próstata más allá de los predictores clínicopatológicos conocidos. Si se confirman, estas variantes genéticas podrían ayudar a aclarar qué tumores pueden progresar y requerir un tratamiento más agresivo en contraste con aquellos que podrían no tener efectos sustanciales sobre la morbilidad o mortalidad.

Rs2292884

LOCUS 2Q37 ALELO DE REFERENCIA A ALELO DE RIESGO G

GENES RELACIONADOS *MLPH*

La expresión de MLPH en tumores primarios de próstata fue significativamente menor en aquellos con la G en comparación con el alelo T y se correlacionó significativamente con la proteína AR. *Un mayor nivel de melanofilina en el tejido prostático de pacientes con un perfil de riesgo de PC favorable señala un efecto supresor del tumor.*

Estos resultados desentrañan un vínculo oculto entre la proteína AR y un riesgo putativo funcional de PC cuya alteración del alelo afecta a la regulación de andrógenos de su gen huésped MLPH.

SNP CROMOSOMA 3

RS2660753

LOCUS 3P12 ALELO DE REFERENCIA C ALELO DE RIESGO T

GENES *DESCONOCIDOS*

Aunque rs2660753 es un polimorfismo de susceptibilidad al cáncer de próstata fuerte, la asociación con otro cáncer relacionado hormonalmente, cáncer ovárico invasivo, no es apoyada por los estudios realizados.

RS7611694

LOCUS 3Q13 ALELO DE REFERENCIA A ALELO DE RIESGO C

GENES RELACIONADOS *SIDT1*

Está relacionado con los genes SIDT1 y se ve que aumenta susceptibilidad.

RS10934853

LOCUS 3Q21 ALELO DE REFERENCIA C ALELO DE RIESGO A

GENES RELACIONADOS *EEFSEC*

			Frequency			
Study population	Cases (*n*)	Controls (*n*)	Cases	Controls	OR (95% CI)	*P* value
Iceland[a]	1,968	35,227	0.295	0.269	1.14 (1.06–1.22)	3.2×10^{-4}
Chicago	1,077	1,003	0.313	0.273	1.21 (1.06–1.39)	4.4×10^{-3}
Finland	2,638	1,716	0.330	0.319	1.05 (0.96–1.15)	0.27
The Netherlands	1,084	1,827	0.306	0.286	1.10 (0.98–1.24)	0.10
Nashville	596	687	0.283	0.270	1.07 (0.90–1.27)	0.47
Spain	811	1,605	0.306	0.314	0.96 (0.84–1.09)	0.54
ACS[b]	1,758	1,775	0.300	0.258	1.25 (1.12–1.39)	4.3×10^{-5}
ATBC[b]	928	921	0.309	0.319	0.96 (0.84–1.10)	0.59
FPCC[b]	654	657	0.291	0.272	1.09 (0.92–1.29)	0.34
HPFS[b]	595	609	0.313	0.278	1.18 (0.99–1.40)	0.070
PLCO[b]	1,167	1,093	0.308	0.266	1.23 (1.08–1.41)	2.5×10^{-3}
CAPS[c]	498	494	0.329	0.288	1.21 (1.00–1.46)	0.045
All combined[d]	**13,774**	**47,614**	**–**	**0.284**	**1.12 (1.08–1.16)**	**2.9×10^{-10}**
P_{het}					**0.039**	
I^2					**46.4**	

En distintos países europeos varios estudios demuestran ese aumento en la susceptibilidad en el desarrollo del PC.

SNP CROMOSOMA 4

RS17021918

LOCUS 4Q22; ALELO DE REFERENCIA C; ALELO DE RIESGO T

GENES RELACIONADOS *PDLIM5*

No hay diferencias estadísticamente significativas en la distribución de frecuencias de los alelos de riesgo y genotipos de PDLIM5, SLC22A3 y NKX3-1 entre los grupos de casos y control (P > 0,05), ni los tres loci gen se asociaron significativamente con la edad, la puntuación de Gleason, el nivel de PSA y el grado patológico de los pacientes PCa (CP < 0,05). El análisis de MDR no mostró interacción entre PDLIM5 y NKX3-1, pero el análisis de diagramas de árboles reveló una *posible acción sinérgica de los loci del polimorfismo.*

RS12500426

LOCUS 4Q22; ALELO DE REFERENCIA C; ALELO DE RIESGO A

GENES RELACIONADOS *PDLIM5*

Los estudios de asociación de todo el genoma (GWAS) han identificado múltiples polimorfismos de nucleótidos individuales (SNP) asociados con el riesgo de cáncer de próstata. Sin embargo, *se desconoce si estas asociaciones pueden replicarse constantemente, variar con la agresividad de la enfermedad (etapa y grado tumoral) y/o interactuar con factores de riesgo potenciales no genéticos u otros SNP*. Ningún SNP mostró asociaciones diferenciales según la etapa o grado de la enfermedad. No observamos ninguna modificación de efecto por parte de SNP para la asociación con la edad en el diagnóstico, antecedentes familiares de cáncer de próstata, diabetes, IMC, altura, tabaquismo o ingesta de alcohol. Además, no encontramos evidencia de interacciones SNP-SNP en pareja. Si bien estos SNP representan nuevos factores de riesgo independientes para el cáncer de próstata, vimos poca evidencia para la modificación del efecto por otros SNP o por los factores ambientales examinados.

SNP CROMOSOMA 5

RS 12653946
LOCUS 5P15
GENES *IRX4*
ALELO DE REFERENCIA C
ALELO DE RIESGO T

Se ha demostrado una *asociación estadística fuerte con el riesgo de desarrollo del Cáncer de Próstata con este SNP concreto en más de 12 estudios publicados en la literatura médica internacional*. En los estudios realizados se han testado diversas poblaciones, razas, etc. confirmando en todos ellos la clara relación con el ulterior desarrollo de dicha tumoración.
Estos genes IRX4 también se han relacionado con el desarrollo de la *miocardiopatía hipertrófica* con aumentos claros del septo interventricular y con la *tetralogía de Fallot*.

RS4001681
LOCUS 5P15CL
GENES *CLPTM1*
ALELO DE REFERENCIA G
ALELO DE RIESGO A

Una proteína transmembrana 1 labial e *paladar hendido* es una proteína que en los humanos está codificada por *genes CLPTM1*. Pertenece a una familia de varias secuencias de proteína transmembrana 1 labial e fisura eucariota.
Está asociado con el *acortamiento de los telómeros* y con el desarrollo de *Cáncer de Pulmón, Melanoma y también con el Prostático que* ahora nos ocupa.

SNP CROMOSOMA 6

RS130067

LOCUS 6P21

GENES *CCHCR1*

ALELO DE REFERENCIA T; ALELO DE RIESGO G

Variantes genéticas comunes asociadas con la enfermedad en estudios de asociación genómico son mutuamente excluyentes para el *cáncer de próstata* y la *artritis reumatoide*. Relacionado con tumores de cabeza y cuello; psoriasis y múltiples estudios con vinculación con el cáncer de próstata.

RS1983891

LOCUS 6P21

GENES *FOXP4*

ALELO DE REFERENCIA C; ALELO DE RIESGO T

La *proteína de la caja de cabeza* es una proteína que en los seres humanos está codificada por el *gen FOXP4*. Los factores de transcripción de la caja de cabezas de bifurcación desempeñan un papel importante en la regulación de la transcripción de genes específicos del tipo de tejido y células durante el desarrollo y la edad adulta. Muchos miembros de la familia de genes de la caja de cabezas de bifurcación, tienen múltiples papeles en la *oncogénesis de mamíferos*. Este gen puede desempeñar un protagonismo en el desarrollo de *tumores del riñón y la laringe*.

RS3096702

LOCUS 6P21

GENES *NOTCH4*

ALELO DE REFERENCIA G; ALELO DE RIESGO A

RS339331

LOCUS 6Q22

GENES *RFX6*

ALELO DE REFERENCIA C; ALELO DE RIESGO T

RS9364554

LOCUS 6Q25

GENES SLC22A3

ALELO DE REFERENCIA C; ALELO DE RIESGO T

SNP CROMOSOMA 7

RS10486567

LOCUS 7P15

GEN JAZF1

ALELO DE REFERENCIA A; ALELO DE RIESGO G

Banda 7p15.2-15.1

Inicio 27,830,573 bp

Final 28,180,795 bp

A nivel celular forma parte de un complejo represor transcripcional implicado en el metabolismo de los lípidos, en la regulación negativa de la transcripción mediada por una RNA polimerasa II y otras regulaciones de la transcripción.

El gen JAZF1 codifica una proteína nuclear con tres dedos de Zinc cuya función es como "represor transcripcional" y químicamente son del tipo C2H2.

Los tumores endometriales son la consecuencia de las alteraciones cromosómicas relacionadas con este gen.

El SNP RS104866567 representa el más importante marcador de riesgo de Cáncer de Próstata incluido dentro del gen JAZF1 en individuos de raza Europea (Ludmila P.Olsson, 2010).

RS 12155172

LOCUS 7P21

GEN SP8

ALELO DE REFERENCIA A

ALELO DE RIESGO G

El gen SP8 codifica una proteína llamada también SP8 cuya función biológica es un regulador de la transcripción de la familia SP. En ratones ha demostrado ser esencial para el normal desarrollo de las extremidades. Se han encontrado dos variantes de transcripción que codifican diferentes isoformas para este gen.

En hombres de ascendencia africana, tres SNP: rs1512268 en NKX3-1 (8p21.2), *rs12155172* en intergénico (7p15.3) y rs10486567 en JAZF1 (7p15.2) se asociaron positivamente con la enfermedad metastásica en el análisis multivariante (Kolawole, 2012).

RS 6465657

LOCUS 7Q21

GEN LMTK2

ALELO DE REFERENCIA T

ALELO DE RIESGO C

El gen LMTK2 codifica la tirosina quinasa 2 de Lemur que es una enzima que, en los seres humanos, está relacionada con tal gen. La función de esta enzima es un tipo de tirosina quinasa y proteína quinasa. Contiene hélices transmembrana en extremo N-terminal y una cola citoplasmática en el extremo C-terminal con actividad serina/ treonina/ tirosina quinasa. Esta proteína interactúa con el RECEPTOR DE LOS ANDRÓGENOS, el inhibidor 2 (inh2), la proteína fosfatasa 1(pp1C), la P35 y la miosina VI. Los estudios con ratones sugieren un papel esencial en la ESPERMIOGENESIS. Participa en la señalización del factor de crecimiento nervioso NGF-TKA y también tiene una importante misión en el tráfico a través de la membrana endosomal. La cantidad o naturaleza de las transcripciones de ARNm expresados a partir de los genes candidatos LMTK2, HNF1B Y MSMB está alterada en el Cáncer de Próstata y proporciona más evidencia de un papel importante de estos genes en la patogenia de esta enfermedad. Las alteraciones en el uso de isoformas resaltan la importancia potencial del procesamiento alternativo del ARNm y la moderación de la estabilidad del ARNm como mecanismos de la enfermedad potencialmente importantes.

SNP CROMOSOMA 8

RS1512268

LOCUS 8P21

GEN NKX3.1

ALELO DE REFERENCIA G

ALELO DE RIESGO A

NKX3.1 juega un papel esencial en el desarrollo normal de la próstata. La pérdida de función de NKx3.1 conduce a defectos en la secreción de proteínas prostáticas así como de la morfogénesis ductal. La pérdida de función también conduce a la carcinogénesis prostática. En humanos contiene 4 exones. Codifican 234 aminoácidos alcanzando un peso molecular de 35 a 38 KD. En el año 2000 el gen NKX3.1 fue obtenido íntegramente de una librería de ADN. Hay identificadas varias variantes con alteraciones relacionadas. Es bien conocida su relación con la caja de genes de regulación androgénica. NKX3.1 es un supresor de tumor. La pérdida del mismo contribuye a la carcinogénesis del CP. NKX3.1 es regulado por los andrógenos en células prostáticas normales. También está claro que desarrolla un papel fundamental en el desarrollo, diferenciación y reprogramación de STEM CELL en el tejido prostático (Antao et al, 2021)

Un nuevo algoritmo para estudios de asociación multivariante de genoma amplio basados en Extreme Learning Machine y Differential Evolution

1. Introducción

Desde la finalización del Proyecto Genoma Humano [1] y el Proyecto Internacional HapMap [2], la comunidad científica sabe bien que existe un vínculo claro entre algunos genes y ciertos rasgos y enfermedades. Los estudios de asociación de todo el genoma (GWAS) [3] son una metodología poderosa que ha demostrado su importancia en el análisis de problemas genómicos complejos.

Aunque los estudios GWAS presentan ciertas limitaciones, principalmente en lo que respecta a la necesidad de hacer uso de un gran tamaño muestral para realizar el estudio [4–6] y tener una estratificación poblacional [7], así como la alta probabilidad de falsos positivos [8,9] que deben gestionarse con la ayuda de las técnicas estadísticas, el uso de GWAS ha permitido a los investigadores producir algunos resultados científicos sobresalientes.

Los primeros GWAS se remontan a los años 2005 y 2006 [10,11]. Esos primeros estudios fueron de gran interés en ese momento, ya que encontraron variantes comunes asociadas a la degeneración macular relacionada con la edad. El hecho de que este tipo de estudios puedan buscar relaciones genéticas de forma multivariante es uno de los principales puntos a su favor. Esto significa que un GWAS es capaz de tratar con un gran número de SNP simultáneamente y buscar cualquier relación entre ellos y un rasgo particular que se está analizando [12]. También es destacable que GWAS no requiere ninguna hipótesis a priori, ya que solo tienen en cuenta las relaciones de los SNP que se pueden colocar en cualquier gen con el rasgo sin necesidad de estudiar una lista previa de loci [13].

Mientras que en el primer análisis GWAS realizado, el vínculo entre los SNP y los fenotipos se consideró de manera univariada [14], los estudios realizados en los últimos años hicieron uso de metodologías multivariantes [15] que, en algunos casos, incluyen técnicas de aprendizaje automático [12]. Investigaciones anteriores han empleado bosques aleatorios, y este es solo un ejemplo de metodologías de

aprendizaje automático que se utilizan en GWAS en los últimos tiempos [16-18]. En un caso particular, [16], se utilizó un método que se basó en bosques aleatorios diseñados con el propósito expreso de interpretar datos genómicos desequilibrados. Otro trabajo [17] desarrolló un algoritmo llamado Reg-SNPs-intron con la ayuda de clasificadores de bosques aleatorios, mientras que en la investigación realizada por Roshan et al. [18], se empleó bosque aleatorio para estudiar el número de variantes casuales y regiones asociadas identificadas por los mejores SNP, y los resultados obtenidos se compararon con otras metodologías.

Además, la metodología de aumento de gradiente ya se aplicó en el marco de GWAS [19,20]. Trabajos anteriores [19] utilizaron modelos de aumento de gradiente con el objetivo de diferenciar los genes de la enfermedad inflamatoria intestinal (EII) de los genes no IBD mediante el uso de información de datos de expresión y anotaciones de genes. En otros casos del uso del aumento de gradiente [20], su objetivo es analizar los loci genéticos para simplificar la tarea de interpretar estudios genéticos a gran escala, manteniendo su carácter inherentemente imparcial.

También se emplearon máquinas de vectores de apoyo (SVM) para clasificar los genes que causan enfermedades inflamatorias intestinales y los que no [19].

Otro artículo [21] aplicó modelos de clasificación SVM para evaluar el potencial del uso de datos GWAS para predecir la duloxetina, y la investigación adicional [12] los empleó en un algoritmo híbrido que combinó SVM con algoritmos genéticos para determinar qué vías tendrían una influencia en el cáncer colorrectal. Además, se emplearon redes neuronales profundas en el aprendizaje profundo para predecir el efecto de las variantes genómicas en la expresión específica del tejido [22] y el algoritmo genético en el análisis GWAS que hace uso de algoritmos híbridos combinados con SVM [12].

Este estudio se realizó para proponer y validar una nueva metodología basada en técnicas de machine learning que puedan aplicarse en estudios GWAS y que puedan mejorar de cierta manera las actualmente disponibles. Más específicamente, la metodología propuesta puede encontrar los SNP más relevantes en cada vía que conduzcan a determinar si un individuo tiene un rasgo determinado o padece una enfermedad en particular. Más específicamente, el método presentado en este trabajo se basa en la Differentialevolution y las máquinas de aprendizaje extremo. El rendimiento de la nueva metodología propuesta se comprobó con la ayuda de una base de datos GWAS, que ayudó a supervisar el rendimiento

de esta nueva metodología propuesta, al igual que una serie de vías bien conocidas, todo lo cual permitió realizar una comparación de los resultados con los obtenidos en algunos de nuestros trabajos anteriores [12].

2. Materiales y métodos

2.1. Differentialevolution

La Differentialevolution (DE) es un algoritmo metaheurístico que fue propuesto por Storn y Price [23], mediante el cual el uso de operadores efectúa una inicialización aleatoria de la población que luego evoluciona para producir descendencia de prueba. Estos operadores también se emplean en métodos, como el cruce y la mutación. Además, DE utiliza un operador de selección que admite a esta descendencia en la población o la descarta de otra manera, dependiendo de los valores de su función objetiva.

Como se mencionó en el párrafo anterior, DE utiliza una población de vectores elegidos al azar como puntos de partida. Esa población es la primera empleada por el algoritmo, mientras que las siguientes se crean con la ayuda de funciones que modifican los puntos iniciales. DE perturba los vectores que componen una determinada población realizando la diferencia escalonada de dos vectores seleccionados aleatoriamente de dicha población. Para producir el nuevo vector, el resultado se agrega a otro miembro vectorial de la misma población. El procedimiento se repite hasta que todos los miembros de una población han competido contra el vector recién generado.$v_0 v_0$

El tema de la optimización se considera uno de minimización en el que el vector elegido para la siguiente población es aquel cuyo valor de función objetivo es el más bajo. Después de que se prueben todos los vectores de prueba, los sobrevivientes son los miembros de la próxima generación y serán responsables de crear nueva descendencia.

El algoritmo DE requiere una estructura de población. El conjunto de individuos de la -ésima generación está representado por , mientras que cada vector de esa generación puede ser representado por , donde significa que es el -ésimo vector de la -ésima generación.$xP_x v_{ix} iix$

Antes de la inicialización aleatoria de la población, los límites superior e inferior para cada variable deben fijarse para asegurarse de que todos los componentes (variables) de todos los miembros (vectores) de la población

estén dentro de ambos límites. Una vez inicializada la población del algoritmo DE, el algoritmo muta y recombina todos los vectores que forman parte de la población. Como se dijo anteriormente, este proceso se basa en la reescala, resta y suma de acuerdo con la siguiente ecuación [24] que se emplea para obtener cada uno de los nuevos vectores:

$$V_n = v_{r0x} + F(v_{r1x} - v_{r2x}),$$

Dónde
V_n es el nuevo vector que se crea, haciendo uso de dos miembros de la -ª generación.x
F es un factor de escala mayor que 1 sin límite superior.
v_{r0x} es el vector -ésimo de la -ª generación.$r0x$
v_{r1x} es el vector -ésimo de la -ª generación.$r1x$
v_{r2x} es el vector -ésimo de la -ª generación.$r2x$

DE también hace uso del crossover uniforme. Consiste en tomar dos vectores de la misma población y copiar una cierta proporción de componentes de uno a otro con cierta probabilidad. En la selección del algoritmo DE, también se incluye el operador. Funciona de la siguiente manera: si el vector de prueba tiene un valor más bajo que el vector objetivo (es decir, el vector -ésimo de la -ª generación), reemplaza al vector objetivo en la siguiente generación; de lo contrario, el objetivo conserva su lugar en la población.$pV_nV_{ix}ix$

Una vez creada la nueva población, el proceso se repite hasta localizar el óptimo, o se alcanza un criterio de terminación preespecificado. El Apéndice A contiene la Tabla A1 que describe el pseudocódigo del algoritmo DE.

2.2. Máquinas de aprendizaje extremo

Cuando se trata de problemas de regresión y clasificación, la máquina de aprendizaje extremo (ELM) ha demostrado ser un algoritmo de aprendizaje altamente práctico [25]. Un punto importante a su favor es que puede aprender de los datos mucho más rápidamente que otras metodologías de aprendizaje automático [25]. No es necesario ajustar los parámetros de la capa oculta iterativamente, y es posible calcular los pesos de salida empleando la metodología de optimización de mínimos cuadrados [26,27].

En el caso de la presente investigación, el ELM se utiliza para la clasificación. Más específicamente, se empleará una máquina de aprendizaje extremo regularizada (RELM) [28]. Una de las principales ventajas de RELM es que no solo es capaz de obtener una mejor

generalización que reduce el error de entrenamiento, sino que también maximiza la distancia de borde.

RELM puede considerar riesgos empíricos y estructurales al mismo tiempo [29]. El modelo matemático RELM se puede expresar como el siguiente problema de optimización [29]:

$$\min\left(\frac{1}{2}\|\beta\|^2 + \frac{\gamma}{2}\|\varepsilon\|^2\right)^2, \quad (1)$$

Sujeto a

$$\sum_{i=1}^{N} \beta_{ig}\left(a_i x_j + b_i\right) - t_j = \varepsilon_j \quad (2)$$

Dónde
β es un parámetro empleado para suavizar la función de costo.
γ es la proporción de los dos tipos de parámetros de riesgo.
ε representa las diferencias entre los vectores de entidad de referencia y los vectores de entidad generados por la capa oculta del RELM.

Las ecuaciones presentadas anteriormente se pueden convertir en un problema extremo incondicional con la ayuda de una función de Lagrange [30].

2.3. El algoritmo propuesto

El algoritmo presentado en este artículo emplea DE y ELM. El objetivo de este nuevo desarrollo es encontrar un cierto subconjunto de SNPs pertenecientes a una vía previamente definida que sería capaz de realizar una clasificación de individuos en casos y controles con cierta precisión. El diagrama de flujo de esta nueva metodología se presenta en la Figura 1. Aunque las herramientas de aprendizaje automático empleadas son diferentes, los fundamentos generales del algoritmo son similares a los propuestos en investigaciones anteriores [12] que hicieron uso de algoritmos genéticos y máquinas de vectores de soporte.

El algoritmo primero toma todos los SNP relacionados con la ruta que se está considerando. Cabe señalar que cada vez que se ejecuta el algoritmo, solo utiliza aquellos SNP correspondientes a esa vía. Como consecuencia, solo aquellos SNP pertenecientes a este subconjunto terminarán siendo miembros del espacio de búsqueda utilizado por el algoritmo DE. Todas las soluciones candidatas se consideran un argumento, tomando la forma de un vector de números reales y produce un valor que indica la idoneidad de la solución candidata bajo análisis como salida. Tenga en cuenta que en el

caso del presente problema, el vector de números reales es una cadena de "1s" y "0s" que indica qué SNP de los que están en el subconjunto participarán en el modelo ELM empleado como función de aptitud del ELM. En nuestro caso, "1" indica que el SNP participará en el modelo ELM, mientras que "0" significa que no lo hará. También hay que mencionar que para cada miembro de la población, se capacitará un modelo ELM diferente.

En la primera generación de la población, solo habrá un SNP activo en cada uno de sus miembros. Esto significa que la población inicial se construye como si se eligieran filas aleatorias de una matriz de identidad cuadrada de rango igual al número de SNP de la vía en estudio como elementos de la población inicial. En las siguientes generaciones, estos miembros de la población evolucionarán haciendo uso de las reglas de DE. La evolución se realiza de tal manera que solo aquellos miembros de la población con un mayor valor de la función de aptitud física se mantendrán y emplearán para crear la próxima generación de individuos. Lo que esto significa es que en este algoritmo, las variables que están activas en el DE se emplean como información de entrada para el modelo ELM. El siguiente paso consiste en evaluar el rendimiento del modelo ELM.

Tenga en cuenta que la función de aptitud empleada en este algoritmo calcula el área bajo la curva ROC que obtienen los individuos clasificados de acuerdo con los resultados del modelo ELM que hace uso de las variables activas. Al igual que en investigaciones anteriores [12] realizadas por los autores, no se permite el uso de más de un SNP del mismo gen para prevenir el fenómeno de epitasis [31], lo que en esta implementación del algoritmo significa que el valor ROC obtenido por los miembros de la población con epitasis se reemplaza por 0.

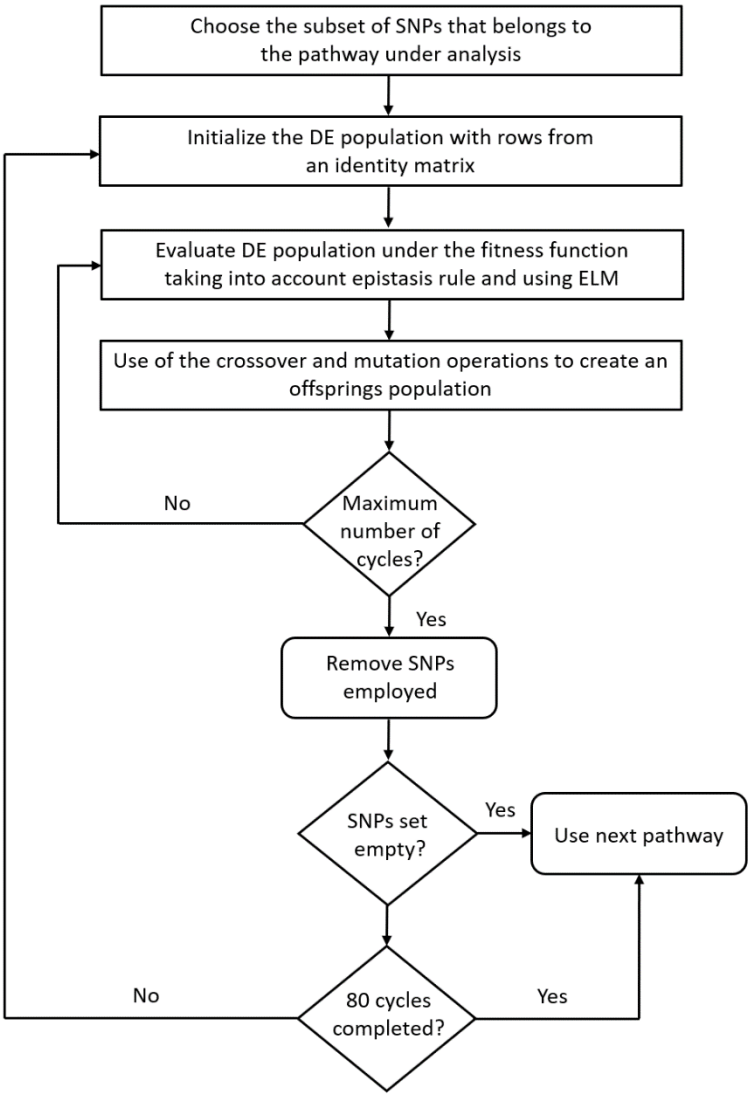


Figura 1. Diagrama de flujo del algoritmo propuesto (DE—Differentialevolution, ELM—máquina de aprendizaje extremo).

El algoritmo se aplica como se muestra en la Figura 1. El primer paso consiste en elegir el subconjunto de SNP que pertenece a la vía bajo análisis. La población DE se inicializa con un conjunto de individuos en los que un SNP está activo. Es por esto que en el diagrama de flujo dice que la población DE se inicializa haciendo uso de las filas de una matriz de identidad. El siguiente paso consiste en evaluar la población DE bajo la función de aptitud, teniendo en cuenta la regla de epitasis y utilizando ELM como algoritmo de clasificación cuyo rendimiento se evalúa mediante el área bajo la curva ROC. Teniendo en cuenta el rendimiento obtenido, se crea una nueva población DE, tomando como punto de partida la población actual. En caso de que se alcance el número máximo de ciclos, se eliminan los SNP ya empleados en el elemento de la población que mejor ha tenido el mejor desempeño. Este proceso se lleva a cabo de

nuevo, y solo se detiene cuando el conjunto de SNP está vacío o cuando se han completado 80 ciclos en total.

El algoritmo propuesto también se aplica mediante la permutación de las etiquetas de casos y controles. Teniendo en cuenta los consejos de la literatura existente, el proceso de permutación se repitió 10.000 veces [32]. Tenga en cuenta que en nuestra investigación anterior [12], debido a limitaciones computacionales, solo se emplearon 1000 permutaciones. Esta opción también es válida, y algunos ejemplos se pueden encontrar en la literatura [33-35], pero en este caso, fue posible realizar 10.000 permutaciones.

2.4. La base de datos

En este trabajo, el conjunto de datos utilizado forma parte del Estudio Transdisciplinario de Cáncer Colorrectal (CORECT), que fue un estudio observacional multicéntrico multicaso control de casos realizado entre septiembre de 2008 y diciembre de 2013, mientras que el subconjunto de información empleada provino del Hospital Universitario León y el Hospital de Bellvitge [12].

Este conjunto de información tuvo un total de 2019 personas, de las cuales 1076 tenían cáncer colorrectal. Para la presente investigación, se emplearon 370.570 SNP distintivos de cada persona. Los casos considerados en esta investigación fueron histológicamente afirmados, y sus edades estuvieron dentro del rango de 20 a 85 años. La elección de los controles se realizó al azar a partir de los registros poblacionales asignados a médicos de familia dentro de la misma área de los hospitales que participaron en esta investigación con la misma edad y sexo, todos habiendo vivido en el área bajo estudio durante al menos 6 meses.

Los sujetos que participaron en este estudio fueron todos voluntarios. La seguridad y privacidad de la investigación fue garantizada a través del protocolo y documento ético aprobado por los Comités de Ética del Estudio. Como no se requería información personal, se eliminó de la base de datos para garantizar la confidencialidad. Todos los expedientes que incluyen información sobre la materia cumplían con la Ley Orgánica 15/1999. Los ficheros empleados en el estudio fueron registrados en la Agencia Española de Protección de Datos con el número de registro 2102672171. Tenga en cuenta también que no solo se obtuvo la aprobación del Comité de Ética del estudio, sino que también se solicitó un doble consentimiento informado a todos los pacientes. Las muestras genéticas se almacenaron en bancos genéticos regionales de las comunidades autónomas de Castilla y León (España) y Cataluña (España).

Es de interés señalar que es posible comparar los resultados de la presente investigación con los obtenidos en trabajos anteriores [12], ya

que se eligieron las mismas vías. Estas son 10 vías que pertenecen a la base de datos KEGG [36–38]. Estas vías se pueden dividir en tres tipos diferentes: vías cuya relación con el cáncer colorrectal ya estaba probada, vías cuya relación con el cáncer colorrectal no es concluyente de acuerdo con la literatura médica actual y, finalmente, vías que según la literatura médica actual es poco probable que tengan una relación con el cáncer colorrectal.

3. Resultados

PathwayName	Tot. SNPs	SNPsEmployed	AUC	AUC Perm.	WinSubsets
Adipocytokinesignalingpathway	752	558	0.542125	0.542302	17.30%
AMPK signalingpathway	1812	508	0.569859	0.555751	89.45%
Apelinsignalingpathway	2525	626	0.578882	0.542724	100%
Colorectalcancerpathway	813	399	0.583862	0.569623	100%
Glucagonsignalingpathway	1707	439	0.563172	0.550494	81.95%
Huntington’sdisease	1980	493	0.552506	0.546617	85.15%
Insulinresistance	1574	489	0.554821	0.558040	30.05%
Insulinsignalingpathway	1215	487	0.557398	0.551704	95.90%
Longevityregulatingpathway	1481	443	0.535514	0.533965	48.85%
Mitochondrialbiogenesis	679	362	0.578295	0.550624	100%

En la presente investigación, el tamaño de la población DE se fijó en 5000,

y el número máximo de iteraciones permitidas fue de 6000. La probabilidad de cruce se fijó en el 50%, y se empleó un factor de ponderación diferencial (F) de 0,8. El algoritmo propuesto se aplicó de la misma manera a todas las vías descritas en la sección de la base de datos.

Este algoritmo primero crea un subconjunto compuesto exclusivamente por SNP pertenecientes a la vía que se está analizando. En la Tabla 1, se puede ver el número de SNP pertenecientes a cada vía. Como muestra esta tabla, la más larga de ellas es la vía de la enfermedad de Huntington con SNP de 1980, mientras que la más corta es la vía de biogénesis mitocondrial con 679.

Tabla 1. Vías en análisis. Número total de SNP por vía (Tot. SNPs), SNPs empleados en las 80 iteraciones por los fenotipos no permutados (SNPs empleados), AUC promedio obtenido en las 80 iteraciones por los fenotipos no permutados (AUC), AUC obtenido por los fenotipos permutados (AUC perm.), porcentaje de valores de AUC no permutados que son superiores al valor máximo de AUC permutado (win subconjuntos).

En total, 5500 individuos conforman la población inicial. Como se mencionó anteriormente en la sección de Materiales y Métodos, solo se encontró un SNP activo en todos estos miembros de la población inicial. Cabe señalar que menos de 2600 SNP están involucrados en cualquiera de las vías que se están analizando, por lo que hay algunos elementos en la población inicial que son los mismos entre sí. Tomando como punto de partida una de esas poblaciones, se crean nuevas generaciones en las que dos o más SNP pueden estar activos simultáneamente. Estos elementos evolucionan, generación tras generación, en busca de la maximización de las AUC. Este proceso se repite hasta llegar a la generación 6000.

El valor de AUC obtenido después de la aplicación del algoritmo a la pathway de adipocitoquina fue de 0,542125, el valor de AUC obtenido para la pathway de AMPK fue de 0,569859, etc. Los valores obtenidos para cada una de las vías bajo análisis se pueden consultar en la columna denominada AUC de la Tabla 1.

En esta investigación, se probó un algoritmo que combina DE y ELM, y en promedio, cada una de las 6000 iteraciones utilizadas tomó 0.73 s. Cuando se completaron las iteraciones, se eliminaron aquellos SNP que se emplearon para entrenar el modelo ELM con el mejor rendimiento, de

modo que el algoritmo repetiría el proceso en busca de aquellos SNP que sean capaces de proporcionar la mejor clasificación pero sin poder hacer uso de aquellos que ya han mostrado el mejor rendimiento de clasificación. Sería posible repetir este proceso mientras la vía tenga SNP disponibles, pero en lugar de repetir el proceso hasta que no hubiera más SNP disponibles, el algoritmo se programó para detener el proceso después de 80 ciclos.

Las razones que llevan a los investigadores a trabajar de esta manera son dos: por un lado, el algoritmo tiene un coste computacional bastante elevado, lo que significa que repetir el bucle mientras todavía hay SNPs disponibles sería un proceso muy lento y, además, ese número de SNPs en las diferentes vías no son los mismos, por lo tanto, detener el proceso antes de quedarse sin SNP facilita la comparación de los resultados.

Una vez finalizada la ejecución del algoritmo, se volvió a ejecutar, haciendo uso de la permutación de las etiquetas de casos y controles pero, conservando el número de individuos en cada una de estas categorías. Tenga en cuenta que el uso de esta metodología es muy común para este tipo de estudios [12].

Los resultados numéricos producidos se presentan en la Tabla 1. Esta tabla contiene la siguiente información sobre todas las vías bajo análisis: cuántos SNP componen el subconjunto de vías, cuántos SNP diferentes de la vía empleada en cualquiera de los modelos mostraron el mejor rendimiento del AUC, qué valor de AUC obtuvo el algoritmo en promedio, el valor del AUC cuando se permutaron los fenotipos y qué porcentaje de valores de AUC no permutados fueron mayores que el valor de AUC permutado más alto que se obtuvo.
En esta investigación, se creó una figura para cada vía bajo análisis. La Figura 2 muestra los valores que se obtuvieron aplicando el algoritmo propuesto a la pathway de adipocitoquina en seis casos diferentes, uno de los cuales es aquel en el que los casos y controles se etiquetaron correctamente como casos y controles, mientras que los otros cinco muestran los resultados obtenidos por cinco permutaciones diferentes. En las seis ejecuciones de algoritmos presentadas, los valores de AUC se ordenan de mayor a menor para facilitar la interpretación de las curvas obtenidas. En el caso de la vía mencionada, la curva que representa datos sin casos y controles de permutaciones no es superior a las curvas obtenidas haciendo uso de casos permutados y etiquetas de controles, ya que están muy cerca unas de otras. Este resultado es confirmado por el diagrama de caja presentado en la Figura 3, donde el valor medio de AUC obtenido para casos y controles correctamente etiquetados es muy similar

a los valores medios alcanzados para las cinco ejecuciones de algoritmos con etiquetas permutadas representadas en la misma figura.

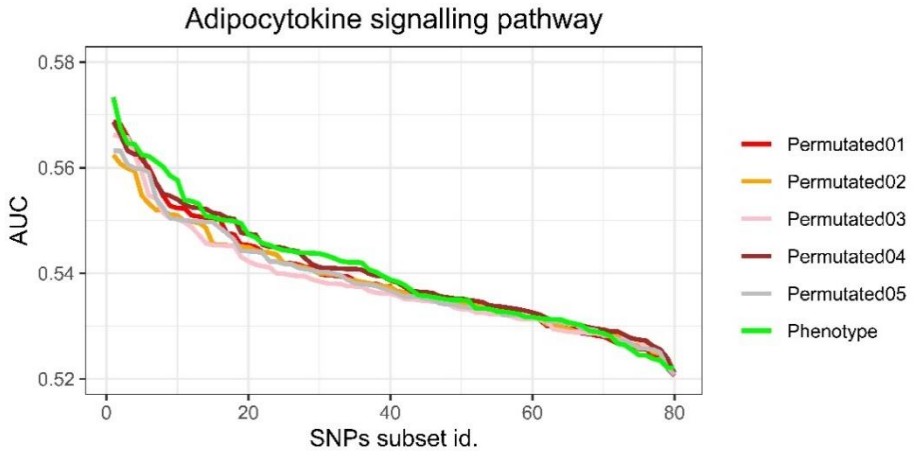


Figura 2. Pathway de Adipocytokine: Valores AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

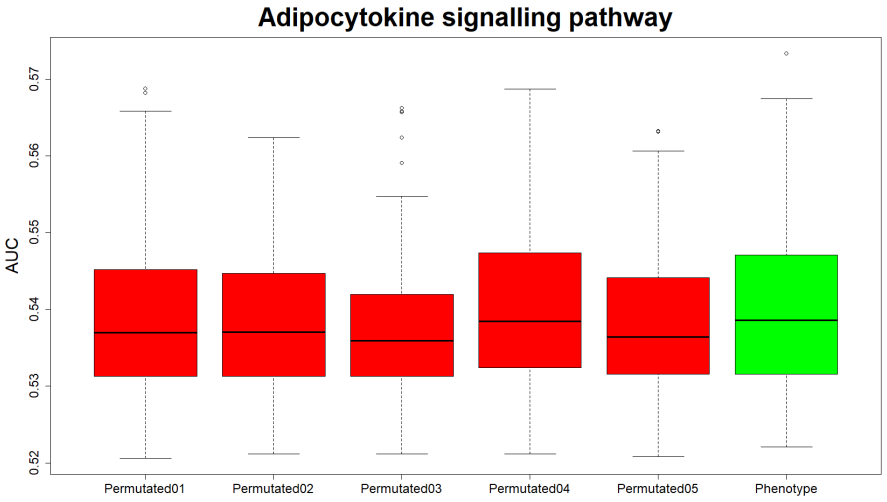


Figure 3. Boxplot of the adipocytokine signaling pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permutated and one with cases and controls correctly labeled. Red color represents permutated cases, green color represents the phenotype case and circles are outliers.

En la Figura 4 se pueden ver los resultados obtenidos para la vía de resistencia a la insulina. La Figura 5 muestra el diagrama de caja de los valores de AUC de resistencia a la insulina de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados,

mientras que la Figura 6 muestra la misma información que la Figura 4 para la vía de longevidad. En ambos casos, y esto también ocurre en la Figura 2, para la curva verde, llamada fenotipo y que se refiere a los resultados producidos por el algoritmo cuando se aplica a los datos con las etiquetas correctamente asignadas a casos y controles, los valores AUC muestran una gran similitud con los obtenidos con etiquetas permutadas. Tenga en cuenta también que en las Figuras 5 y 7 se puede observar el mismo efecto, ya que en ambos casos el valor medio de las AUC para casos y controles no es superior al valor medio de todas las ejecuciones permutadas. Por lo tanto, se debe inferir que estas tres vías no están relacionadas con el cáncer colorrectal.

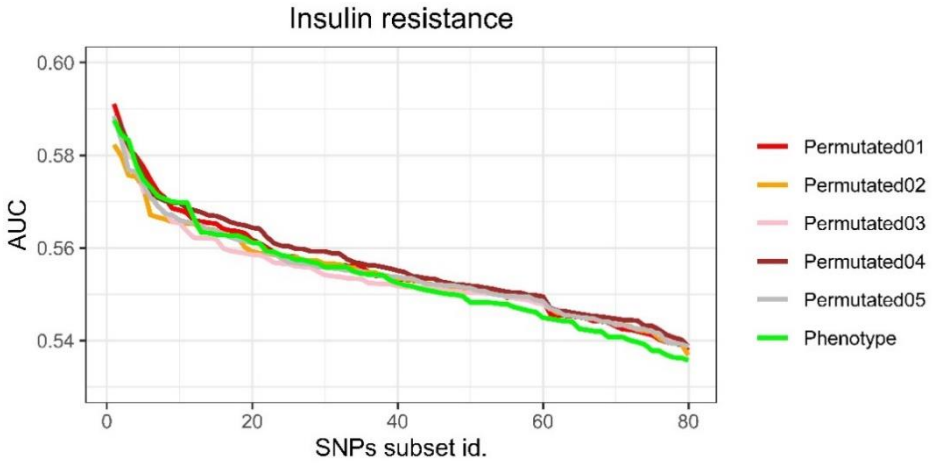


Figura 4. Resistencia a la insulina: Valores de AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente marcados y cinco permutaciones diferentes.

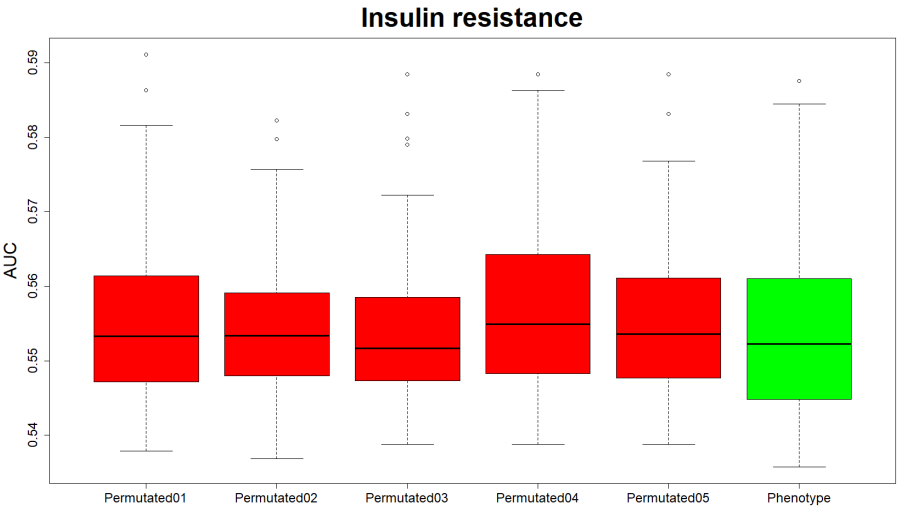


Figura 5. Boxplot de los valores del AUC de resistencia a la insulina de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

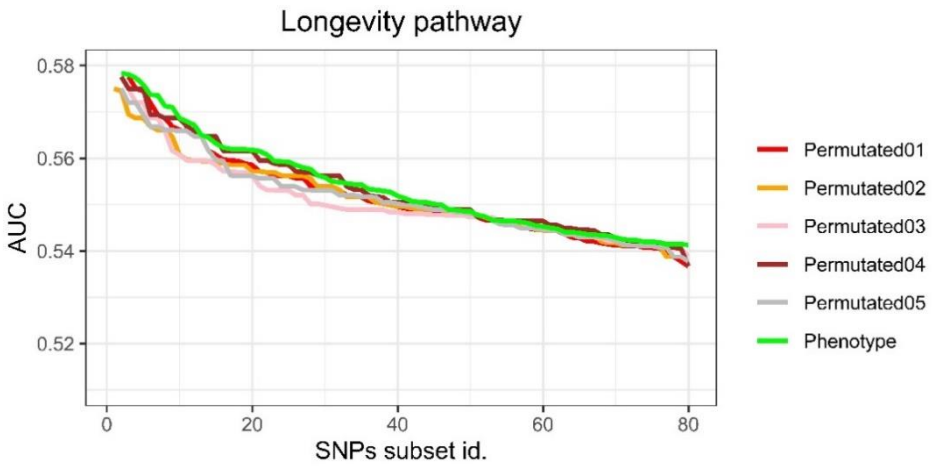


Figura 6. Ruta de longevidad: Valores AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

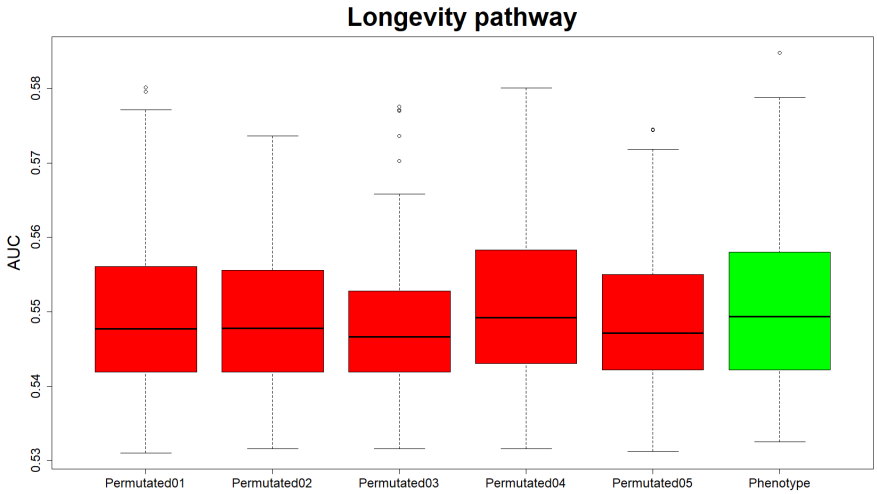


Figura 7. Boxplot de la vía de longevidad AUC valores de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles

correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

Considerando los resultados obtenidos mediante el algoritmo, el valor de AUC obtenido fue en promedio un poco mayor cuando se aplicó el algoritmo a aquellos conjuntos de datos en los que se permutaron las etiquetas de casos y controles que el logrado para casos y controles. En cuanto a la vía de resistencia a la insulina, hubo un mayor valor para los casos y controles, por una diferencia de 0,003219 o 0,58%, mientras que para la vía de longevidad, esta diferencia fue de 0,001549, o 0,29%. Un resumen de estos resultados se puede ver en la Tabla 1.

Las tres vías siguientes en estudio fueron la pathway de la apelina que se representa en las Figuras 8 y 9, la vía de biogénesis mitocondrial que se presenta en las Figuras 10 y 11 y la vía del cáncer colorrectal que se puede observar en las Figuras 12 y 13. En estas tres vías, la curva que representa los valores de AUC obtenidos por el algoritmo cuando se aplican al subconjunto en el que las etiquetas corresponden a casos y control tiene valores indudablemente más altos que los obtenidos cuando se aplicó la permutación. El mismo efecto se observa en sus correspondientes diagramas de caja, ya que el valor medio para la ejecución del algoritmo sin permutación de casos y controles es mayor que las cinco ejecuciones en las que se permutan casos y controles. Incluso en el caso de la pathway de la apelina y la vía de biogénesis mitocondrial, el valor que corresponde al percentil 25 de las ejecuciones sin permutación es superior al percentil 75 de la mayoría de las ejecuciones permutadas. Estos resultados gráficos están en línea con los presentados en la Tabla 1, donde el valor del AUC para la pathway de la apelina fue de 0,578882, en comparación con 0,542724 cuando se permutó. Algo similar ocurrió en el caso de la vía de biogénesis mitocondrial, donde el valor de AUC alcanzado fue de 0,578295, mientras que con las etiquetas permutadas, el valor medio fue de 0,550624. En el caso de la vía del cáncer colorrectal, el valor del AUC fue de 0,583862, mientras que en el de las vías permutadas, el valor medio fue de 0,569623. También es de interés destacar que en estas tres vías, el 100% de los subconjuntos permutados obtuvieron resultados bajo el subconjunto con casos y controles correctamente etiquetados.

En otras palabras, estas tres vías son las más propensas de las 10 bajo análisis en la presente investigación para influir sobre el cáncer colorrectal. Esto se puede notar claramente cuando se comparan sus

cifras con, por ejemplo, las que corresponden a los resultados obtenidos para la pathway de adipocitoquina (Figura 2), la vía de resistencia a la insulina (Figura 4) o la vía de longevidad (Figura 6) o incluso con otras que se describirán posteriormente en la presente sección. Desde el punto de vista del aprendizaje automático, se puede interpretar que estas vías son las que son capaces de separar de forma más clara los casos de los controles, pero desde un punto de vista genético, se traduce como una posible influencia de la vía objeto de estudio sobre el rasgo de interés que en la presente investigación está sufriendo o no de cáncer colorrectal. Una descripción detallada de las posibles razones biológicas de la importancia de las tres vías referidas en el cáncer colorrectal se puede encontrar en la sección de discusión.

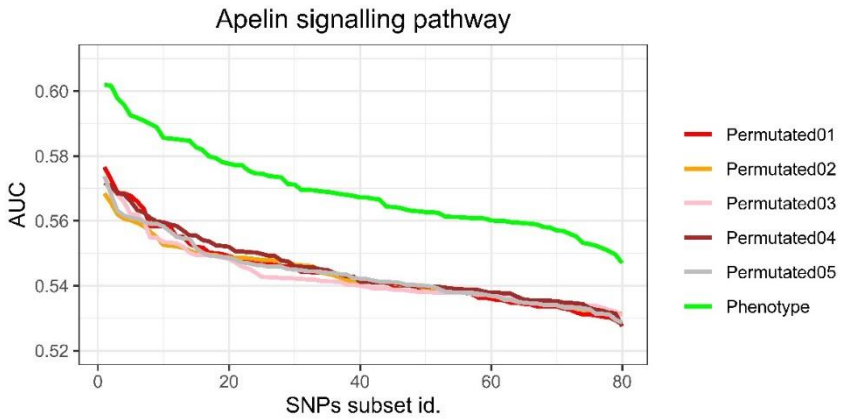


Figura 8. ApelinSignallingpathway: Valores AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

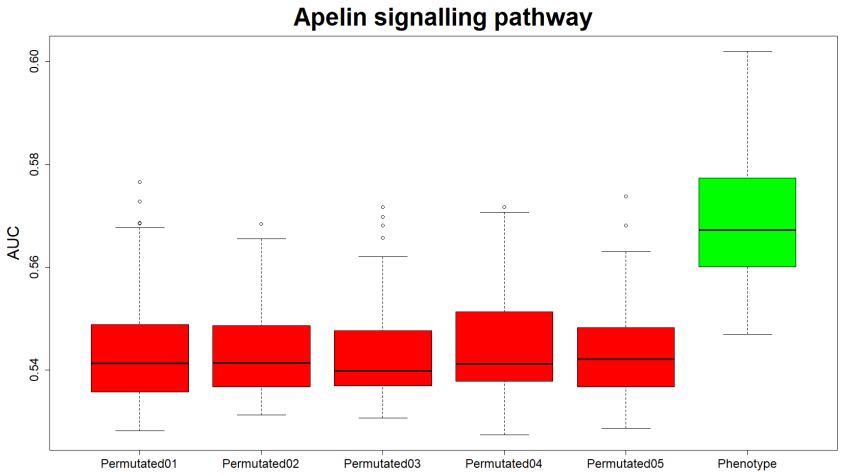


Figura 9. Boxplot de la pathway de apelina Valores AUC de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

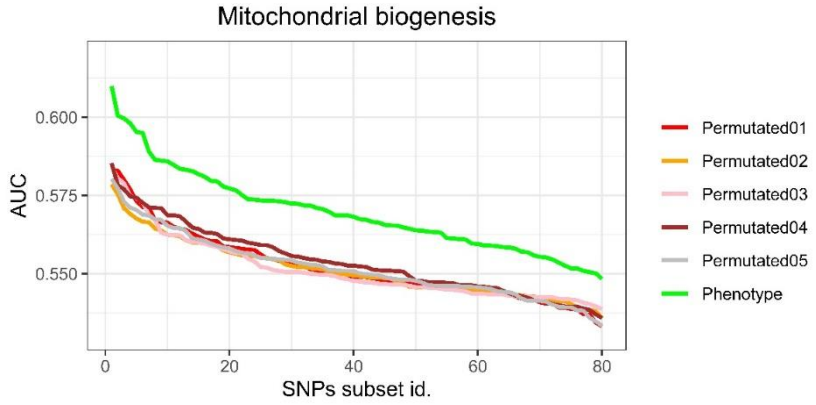


Figura 10. Vía de biogénesis mitocondrial: Valores de AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

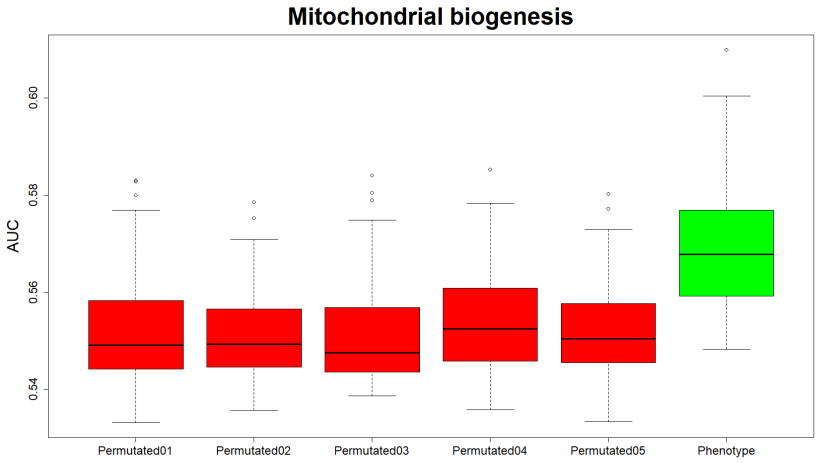


Figura 11. Boxplot de la vía de biogénesis mitocondrial Valores del AUC de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

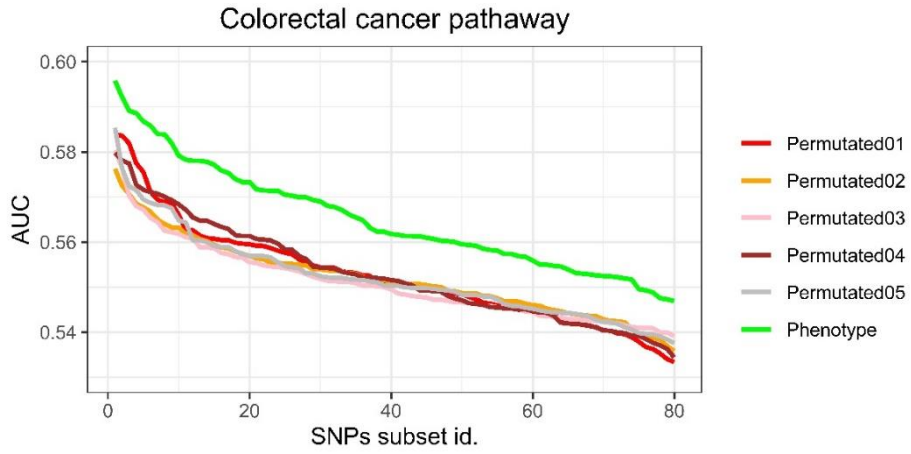


Figura 12. Vía del cáncer colorrectal: Valores de AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

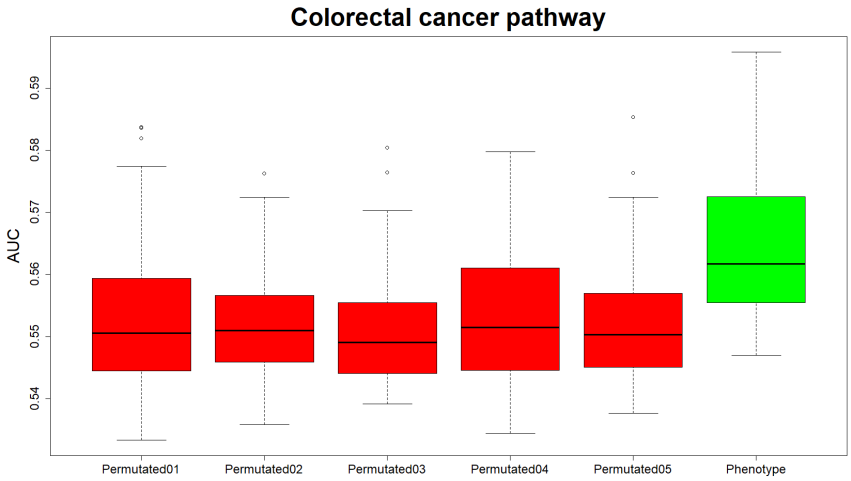


Figura 13. Boxplot de los valores de AUC de la vía del cáncer colorrectal de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

La Figura 14 representa cómo se comporta el algoritmo cuando se aplica a la pathway AMPK. Para esta vía, la curva del fenotipo es cercana a las permutadas. El mismo resultado se observa en la trama de caja de la Figura 15. Los resultados numéricos, presentados en la Tabla 1, donde el valor del AUC de la curva del fenotipo es mayor que el valor obtenido cuando los casos y controles fueron permutados, muestran un resultado ligeramente mejor en el caso de no permutación. En esta figura, las curvas de casos y control tienen un mejor desempeño que el 89,45% de las curvas permutadas. En otras palabras, hay algunos subconjuntos etiquetados aleatoriamente donde el algoritmo logra mejores resultados de AUC que en el caso de casos y controles.

Las figuras 16 y 17 muestran los resultados obtenidos para la pathway del glucagón, mientras que las figuras 18 y 19 hacen lo mismo para la vía de la enfermedad de Huntington. En ambos casos, la curva del fenotipo está muy cerca de las curvas permutadas, aunque su valor medio parece ser un poco más alto. Teniendo en cuenta los valores presentados en la Tabla 1, en ambas vías, el valor AUC de los casos y control es ligeramente superior al de los permutados. En el caso de la pathway del glucagón, el 81,95% de los subconjuntos permutados obtuvieron un valor de AUC inferior a los no

permutados, mientras que en el caso de la vía de la enfermedad de Huntington, estos porcentajes se elevaron al 85,15%.

Finalmente, la Figura 20 muestra los resultados obtenidos cuando el algoritmo se aplica a la pathway de la insulina. En este caso, la curva del fenotipo está cerca de la permutada, pero en la mayoría de los puntos, la primera está por encima de la segunda. En el caso de la trama de caja presentada en la Figura 21, la mediana del valor AUC de la ejecución sin permutación es mayor que las cinco permutaciones mostradas para la misma vía. Tenga en cuenta también que en este caso, el valor del AUC para casos y controles es superior al valor permutado promedio y es superior al 96,5% de los resultados permutados obtenidos.

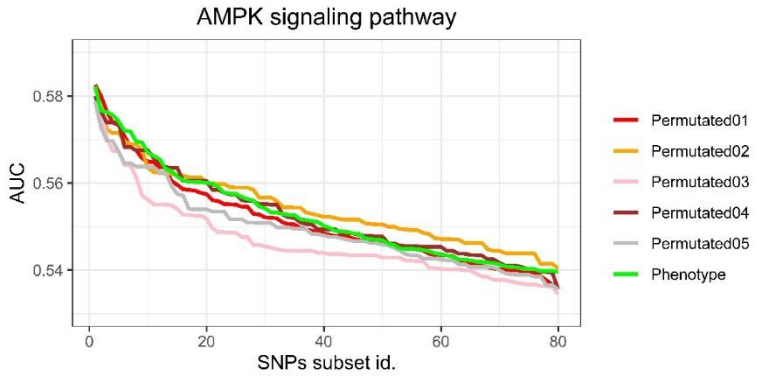


Figura 14. Pathway AMPK: Valores AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

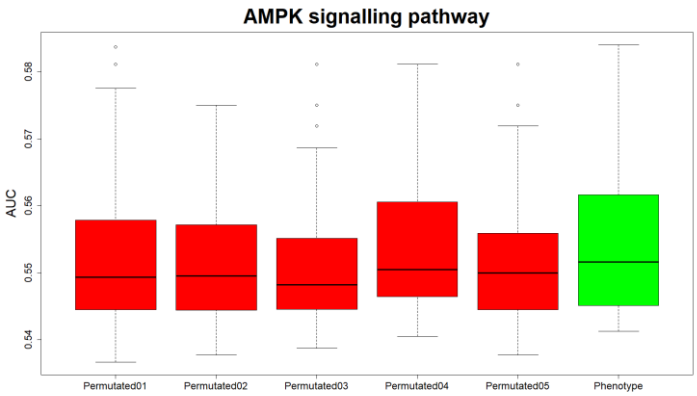


Figura 15. Boxplot de los valores AUC de la pathway AMPK de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

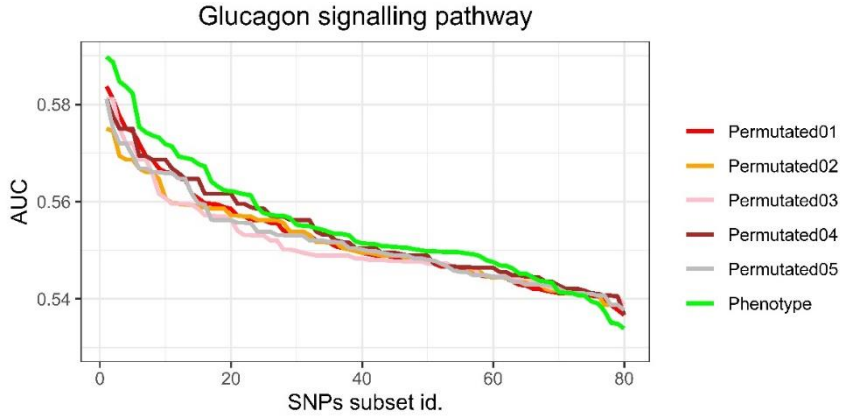


Figura 16. Pathway glucagón: Valores de AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

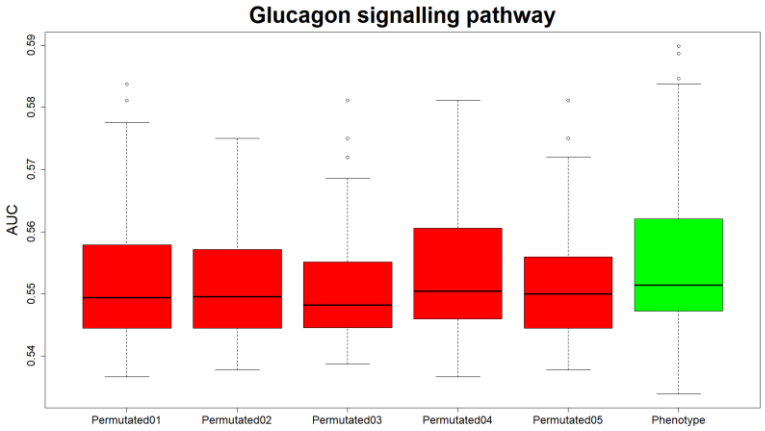


Figura 17. Diagrama de caja de la pathway glucagón Valores AUC de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

Figura 18. Vía de la enfermedad de Huntington: valores de AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

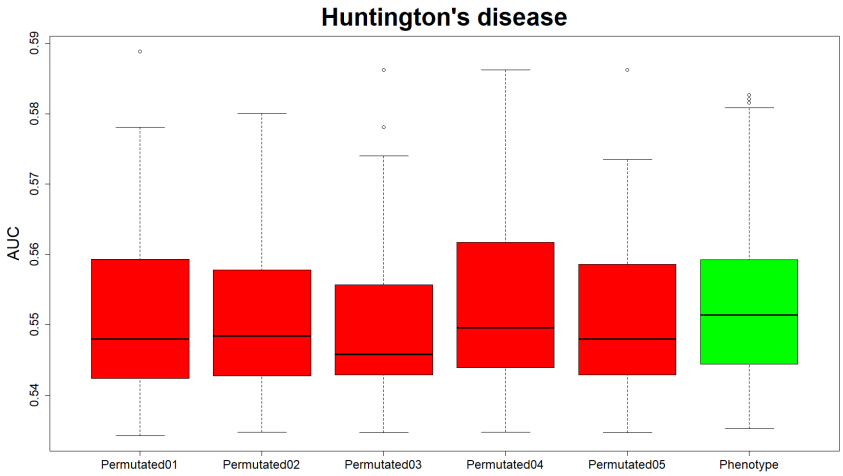


Figura 19. Boxplot de los valores AUC de la vía de la enfermedad de Huntington de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

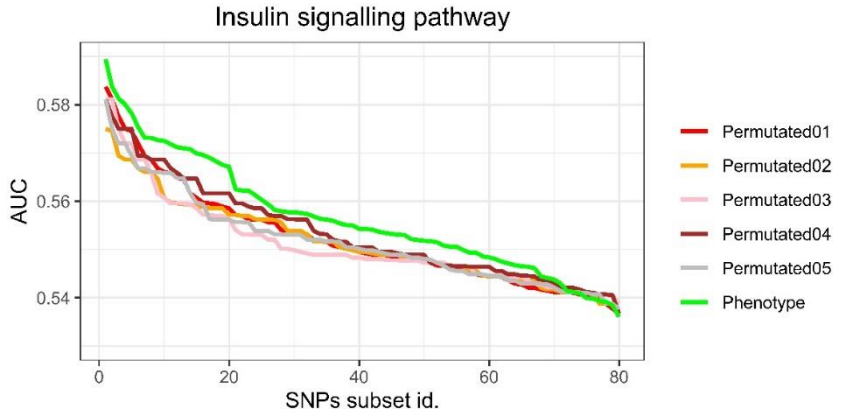


Figura 20. Pathway de insulina: valores de AUC de 80 iteraciones para la ejecución del algoritmo con casos y controles correctamente etiquetados y cinco permutaciones diferentes.

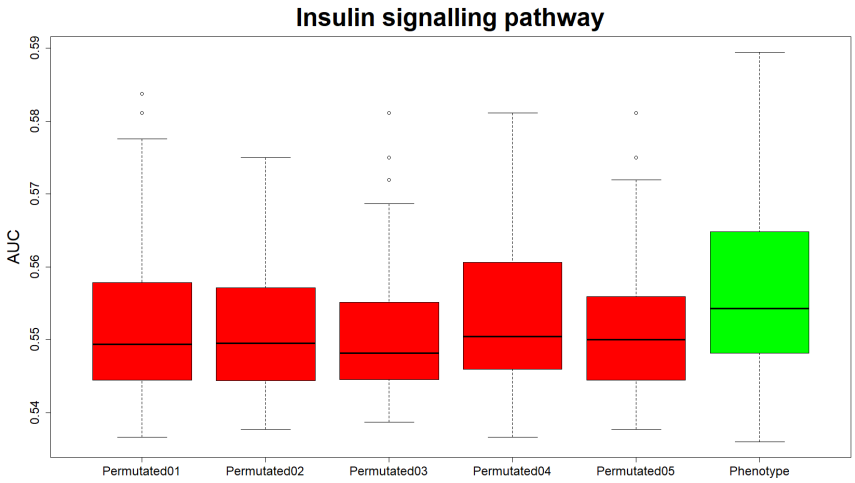


Figure 21. Boxplot of the insulin-signaling pathway AUC values of the 80 iterations of 5 executions

Figura 21. Boxplot de la pathway de insulina. Valores de AUC de las 80 iteraciones de 5 ejecuciones del algoritmo con las etiquetas de casos y controles permutados y una con casos y controles correctamente etiquetados. El color rojo representa los casos permutados, el color verde representa el caso del fenotipo y los círculos son valores atípicos.

4. Discusión

Los resultados obtenidos por el algoritmo propuesto en la presente investigación que hace uso de ELM para la clasificación son bastante similares a los obtenidos en un estudio realizado por los autores en el que se emplearon SVM y algoritmos genéticos [12]. El rendimiento de ELM fue bastante similar al obtenido por SVM, pero el tiempo requerido para el entrenamiento del modelo fue considerablemente menor (tiempos aproximadamente divididos por 47), lo que hace que el uso de ELM sea especialmente conveniente teniendo en cuenta el largo tiempo requerido para el entrenamiento del modelo SVM.

Como ya se dijo en la sección de materiales y métodos, ELM es un algoritmo de aprendizaje basado en redes neuronales de avance que hace uso de una sola capa oculta [39]. Aunque ELM tiene una capacidad de generalización más débil que SVM para una muestra pequeña, puede generalizar, así como SVM para muestras grandes.

En nuestra investigación, como en trabajos anteriores, se encontró que ELM tiene ventajas sobre SVM en la selección de parámetros [40]. Al igual

que SVM, ELM es capaz de minimizar los errores de entrenamiento, así como maximizar el margen de separación [41].

A pesar de que existe una gran cantidad de literatura existente, todavía es necesaria una investigación más profunda sobre la comparación de ELM y SVM. La presente investigación confirma un resultado teórico encontrado en la literatura existente que indica que ELM puede lograr una precisión similar a SVM [40].

Investigaciones anteriores [39] realizaron una comparación experimental de SVM y ELM para diferentes tamaños de muestra de entrenamiento. Esta investigación empleó ocho conjuntos de datos diferentes con muestras que van de 150 a 22, 784 y el número de variables de 4 a 8. Estos resultados sugirieron que SVM tiene el comportamiento de generalización más fuerte. Este hecho es importante en el caso de tamaños de muestra pequeños, pero cuando el tamaño del conjunto de datos de entrenamiento aumenta, la capacidad de generalización del ELM se acerca más al SVM, haciendo que ambos tengan capacidades de clasificación similares para tamaños de muestra grandes. Según investigaciones anteriores [42], DE supera a GA en muchos problemas de optimización, tanto individuales [24] como multiobjetivos [43].

En el caso del presente trabajo, el tamaño muestral seleccionado fue el mismo que en investigaciones anteriores [12] para permitir comparar los resultados. Tenga en cuenta también que en este caso, en lugar de 1000 permutaciones, se realizó un número elegido debido al alto costo computacional establecido en dicha investigación [12], se realizaron 10,000 permutaciones.

Desde nuestro punto de vista, los resultados alcanzados en este trabajo no solo son muy similares a los obtenidos en trabajos anteriores de [12] que hicieron uso de la misma base de datos, sino que también están en línea con otros disponibles en la literatura existente. Una de las principales preocupaciones de los lectores que no están familiarizados con los estudios de GWAS es que los valores de AUC obtenidos son bastante bajos, sin embargo, este es frecuentemente el caso [44] en esta área de investigación.

Teniendo en cuenta estos resultados, concluimos que algunas de las vías propuestas están obviamente relacionadas con el cáncer colorrectal, a saber, la señalización de la apelina, el cáncer colorrectal y la biogénesis mitocondrial. Otros presentan una relación probable, aunque débil, con el cáncer colorrectal, a saber, la señalización de AMPK, la señalización de

glucagón, la enfermedad de Huntington y la señalización de insulina. No se encontró ninguna relación para la señalización de adipocitoquina, la resistencia a la insulina o la regulación de la longevidad.

Con respecto al cáncer colorrectal, investigaciones anteriores ya han destacado la importancia de la pathway de la apelina [45]. La apelina (AP) puede ser potencialmente un objetivo para la terapia contra el cáncer [46,47]. Investigaciones adicionales [48] estudiaron los tejidos tumorales de 56 pacientes con adenocarcinoma colorrectal tratados quirúrgicamente y en ellos se llevó a cabo un análisis de la apelina y su RECEPTOR de ARNm, así como los niveles de expresión de proteínas. Los valores obtenidos se compararon con 27 controles sanos, encontrando que los niveles séricos de apelina y su receptor aumentaron en pacientes con cáncer colorrectal en comparación con los controles. Estos resultados llevan a la conclusión de que la apelina es un factor importante en la progresión del carcinoma colorrectal. Finalmente, un estudio publicado recientemente llegó a la conclusión de que el nivel de apelina y sus receptores está estrechamente relacionado con la regulación de la migración y la invasión de células de cáncer de colon [49]. Se esperaba la presencia de la vía del cáncer colorrectal como una de las vinculadas a pacientes que padecen cáncer colorrectal, y puede considerarse una prueba básica para garantizar el correcto comportamiento del algoritmo.

Como es bien sabido y ya se ha dicho en investigaciones anteriores, las mitocondrias están relacionadas con la génesis del cáncer [12]. De hecho, el desarrollo del cáncer en humanos está estrechamente relacionado con la alteración mitocondrial, el aumento de la producción de radicales libres de óxido mitocondrial y el estrés oxidativo [50].

Existen otras vías en las que se encontró un cierto grado de relación con las etiquetas de casos y controles de cáncer colorrectal. El primero de ellos es la pathway AMPK. En la literatura existente, ahora hay algunos artículos que lo vinculan con el cáncer colorrectal, ya sea considerando que AMPK promueve la supervivencia de las células madre del cáncer colorrectal [51] o mostrando que la expresión de p-AMPK es más frecuente en los controles que en los casos colorrectales [52].

El hecho de que el glucagón aumenta la producción de glucosa en el hígado al aumentar la glucogénesis y la gluconeogénesis al tiempo que reduce la glucólisis de la glucogénesis es ahora ampliamente conocido. La investigación realizada en 2008 [53] descubrió expresiones del receptor de glucagón en líneas celulares de cáncer de colon y en tejido de cáncer de colon obtenido de pacientes. Además, otro estudio que salió el mismo año

informó que el crecimiento de las células de cáncer de colon es promovido por el glucagón a través de la regulación de las vías AMPK y MAPK [54].

En el caso de la enfermedad de Huntington, estudios previos han encontrado [55] que aquellos que sufren de esta enfermedad tienen hasta un 80% menos de cáncer que la población general. La razón es que el gen de la huntingtina mutada (HTT) en aquellos que sufren de la enfermedad de Huntington genera una clase de moléculas pequeñas que son altamente tóxicas para las células cancerosas.

La relación entre la pathway de la insulina y el riesgo de cáncer colorrectal está respaldada por investigaciones anteriores [56] que encontraron que las variaciones genéticas en los genes de las vías de señalización de la insulina pueden afectar el riesgo de cáncer colorrectal. También hay que tener en cuenta que la modificación en los valores individuales de los niveles plasmáticos de insulina debido a la dieta también puede afectar al riesgo de padecer cáncer colorrectal [56]. Finalmente, debe destacarse que no se encontró relación entre las vías de señalización de adipocitoquina, resistencia a la insulina y regulación de la longevidad. Estos resultados están en línea con la falta de resultados sobre estas posibles relaciones encontradas en la literatura.

Aunque el algoritmo presentado en esta investigación ha demostrado una capacidad predictiva satisfactoria, esta capacidad carece de fácil interpretabilidad biológica [57]. Para hacer frente a esta limitación, se requiere el uso de técnicas explicables de inteligencia artificial. Como ya señalaron otros autores [58], la principal ventaja de las técnicas explicables de inteligencia artificial es que integran la interpretabilidad y la transparencia en los modelos de aprendizaje automático [59], lo que, en el caso del problema en estudio, significa que se conocería la relevancia de los diferentes SNP en una determinada vía. Además, hay que destacar que desde un punto de vista biológico, el uso de sistemas explicables de técnicas de inteligencia artificial ayudaría a los profesionales sanitarios a conocer mejor el modelo y a tomar decisiones razonadas.

Finalmente, desde el punto de vista de los investigadores, una de las primeras técnicas explicables de inteligencia artificial que se deben probar con la base de datos de la presente investigación es el bosque aleatorio. El bosque aleatorio es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado de forma independiente y con la misma distribución para cada uno de ellos [60]. Se puede considerar una modificación sustancial del embolsado que

construye una gran colección de árboles no correlacionados y luego los promedia [61].

5. Conclusiones

La investigación descrita en este trabajo presenta un algoritmo novedoso basado en metodologías de aprendizaje automático que ha demostrado dar un buen rendimiento en GWAS. Este trabajo continúa una línea de investigación [12] que hace uso de algoritmos y combina diferentes metodologías de aprendizaje automático. Uno de los principales inconvenientes de estas metodologías es la falta de una explicación biológica simple de los resultados obtenidos, ya que, aunque existen muchas vías cuyas relaciones con la enfermedad y los rasgos son bien conocidas, es difícil encontrar cómo se comporta cada uno de los SNP que forman la vía en el proceso e influye en él. A pesar de ello, es posible explicar la influencia de las diferentes vías sobre el cáncer colorrectal haciendo uso de la literatura disponible.

As was already stated in a previous work [12], in our opinion, due to the current lack of a gold standard multivariate methodology for GWAS, all the algorithms,such as the one presented in this research, should be taken into account in GWAS. Additionally, when compared with the previous algorithm proposed by the authors [12], the fact that the computation time required for this one is about 1 divided by 47 must be considered as one of the main advantages of the algorithm proposed in this research. This result is mainly due to the fast training of the ELM that was already stated in the literature [62].

Como ya se ha dicho en un trabajo anterior [12], en nuestra opinión, debido a la falta actual de una metodología multivariante gold estándar para GWAS, todos los algoritmos, como el presentado en esta investigación, deben tenerse en cuenta en GWAS. Además, en comparación con el algoritmo anterior propuesto por los autores [12], el hecho de que el tiempo de cálculo requerido para este sea de aproximadamente 1 dividido por 47 debe considerarse como una de las principales ventajas del algoritmo propuesto en esta investigación. Este resultado se debe principalmente al rápido entrenamiento del ELM que ya fue declarado en la literatura [62].

Además, como se demostró en la presente investigación, y en línea con trabajos anteriores, el aprendizaje automático es una herramienta valiosa para el análisis GWAS, ya que es capaz de encontrar SNP y los loci de interés. A pesar de ellos, uno de los principales inconvenientes del

machine learning es la falta de una explicación clara de los modelos obtenidos con la mayoría de las metodologías. Este hecho es importante en el campo de GWAS, donde las relaciones entre genes y rasgos son a menudo difíciles de interpretar. Teniendo en cuenta lo mencionado anteriormente, una de nuestras futuras líneas de investigación consistirá en aplicar un algoritmo explicable de aprendizaje automático, como el bosque aleatorio, al problema en estudio en este trabajo. Adicionalmente, y con el fin de contribuir a consolidar el papel de los algoritmos de machine learning en GWAS, los autores también se centrarán en comparar los resultados obtenidos con otras metodologías que son comunes en GWAS y que no pertenecen al campo del machine learning con el fin de cerrar la brecha entre los diferentes enfoques de GWAS.

Apéndice A

La Tabla A1 muestra el algoritmo del algoritmo DE. Este pseudocódigo requiere 4 índices, uno de los cuales es el índice de destino, mientras que los otros tres son los índices vectoriales, llamados , y . Tenga en cuenta que y . Cuando se completa la población, se realiza la selección. En este pseudocódigo, representa el número de elementos en la población.$ir_0r_1r_2r_0 \neq r_1r_1 \neq r_2N_p$

Cuadro A1. Algoritmo del algoritmo de Differentialevolution (DE).

```
// initialize...
do // generate a trial population
{
   for (i=0; i<Np; i++)  // r0!=r1!=r2!=i
   {
      do r0=floor(rand(0,1)*Np); while (r0==i);
      do r1=floor(rand(0,1)*Np); while (r1==r0 or
r1==i);
      do r2=floor(rand(0,1)*Np); while (r2==r1 or
r2==r0 or r2==i);
jrand=floor(D*rand(0,1));
      for (j=0; j<D; j++)  // generate a trial vector
      {
         if (rand (0,1)<=Cr or j==jrand)
        {
u_{j,i} = x_{j,r0} + F * (x_{j,r1} - x_{j,r2});  //check for out-of-
bounds ?
         }
```

```
            else
            {
u_{j,i} = x_{j,i};
            }
        }
    }
    // select the next generation
    for (i=0; i<Np; i++)
    {
        if ( f(u_i) <= f(x_i) ) x_i = u_i;
    }
} while (termination criterion not me);
```

References

1. Venter, J.C. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351.
2. Gibbs, R.A.; Belmont, J.W.; Hardenbol, P.; Willis, T.D.; Yu, F.L.; Yang, H.M.; Ch'ang, L.Y.; Huang, W.; Liu, B.; Shen, Y. The International HapMap Project. *Nature***2003**, *426*, 789–796.
3. Manolio, T.A. Genomewide Association Studies and Assessment of the Risk of Disease. *N. Engl. J. Med*.**2010**, *363*, 166–176.
4. Nishino, J.; Ochi, H.; Kochi, Y.; Tsunoda, T.; Matsui, S. Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. *Front. Genet.***2018**, *9*, 227.
5. Hong, E.P.; Park, J.W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genom. Inform.* **2012**, *10*, 117–122.
6. Ziyatdinov, A.; Kim, J.; Prokopenko, D.; Privé, F.; Laporte, F.; Loh, P.-R.; Kraft, P.; Aschard, H. Estimating the Effective Sample Size in Association Studies of Quantitative Traits. *G3* **2021**, *11*, jkab057.
7. Hellwege, J.N.; Keaton, J.M.; Giri, A.; Gao, X.; Velez Edwards, D.R.; Edwards, T.L. Population Stratification in Genetic Association Studies. *Curr. Protoc. Hum. Genet.***2017**, *95*, 1.22.1–1.22.23.
8. Platt, A.; Vilhjálmsson, B.J.; Nordborg, M. Conditions under Which Genome-Wide Association Studies Will Be Positively Misleading. *Genetics***2010**, *186*, 1045–1052.
9. Shen, X.; Carlborg, O. Beware of Risk for Increased False Positive Rates in Genome-Wide Association Studies for Phenotypic Variability. *Front. Genet.***2013**, *4*, 93.
10. Klein, R.J.; Zeiss, C.; Chew, E.Y.; Tsai, J.Y.; Sackler, R.S.; Haynes, C.; Henning, A.K.; SanGiovanni, J.P.; Mane, S.M.; Mayne, S.T. Complement factor H polymorphism in age-related macular degeneration. *Science***2005**, *308*, 385–389.

11. DeWan, A.; Liu, M.; Hartman, S.; Zhang, S.S.; Liu, D.T.; Zhao, C.; Tam, P.O.; Chan, W.M.; Lam, D.S.; Snyder, M. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science***2006**, *314*, 989–992.
12. Díez Díaz, F.; Sánchez Lasheras, F.; Moreno, V.; Moratalla-Navarro, F.; Molina de la Torre, A.J.; Martín Sánchez, V. GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. *Mathematics***2021**, *9*, 654.
13. Ziegler, A.; Ghosh, S.; Dyer, T.D.; Blangero, J.; Maccluer, J.; Almasy, L. Introduction to genetic analysis workshop 17 summaries. *Gen. Epidemiol.***2011**, *35*, S1–S4.
14. Lippert, C.; Listgarten, J.; Davidson, R.I.; Baxter, S.; Poon, H.; Cadie, C.M.; Heckerman, D. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Sci. Rep.***2013**, *3*, 1099.
15. Ning, C.; Wang, D.; Zhou, L.; Wei, J.; Liu, Y.; Kang, H.; Zhang, S.; Zhou, X.; Xu, S.; Liu, J.F. Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. *Bioinformatics***2019**, *35*, 4879–4885.
16. Schubach, M.; Re, M.; Robinson, P.N.; Valentini, G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Sci. Rep.***2017**, *7*, 2959.
17. Lin, H.; Hargreaves, K.A.; Li, R.; Reiter, J.L.; Wang, Y.; Mort, M.; Cooper, D.N.; Zhou, Y.; Zhang, C.; Eadon, M.T. RegSNPs-Intron: A Computational Framework for Predicting Pathogenic Impact of Intronic Single Nucleotide Variants. *Genome Biol.***2019**, *20*, 254.
18. Roshan, U.; Chikkagoudar, S.; Wei, Z.; Wang, K.; Hakonarson, H. Ranking Causal Variants and Associated Regions in Genome-Wide Association Studies by the Support Vector Machine and Random Forest. *Nucleic. Acids Res.***2011**, *39*, e62.
19. Isakov, O.; Dotan, I.; Ben-Shachar, S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. *Inflamm. Bowel Dis.***2017**, *23*, 1516–1523.
20. Deo, R.C.; Musso, G.; Tasan, M.; Tang, P.; Poon, A.; Yuan, C.; Felix, J.F.; Vasan, R.S.; Beroukhim, R.; De Marco, T. Prioritizing Causal Disease Genes Using Unbiased Genomic Features. *Genome Biol.***2014**, *15*, 534.
21. Maciukiewicz, M.; Marshe, V.S.; Hauschild, A.-C.; Foster, J.A.; Rotzinger, S.; Kennedy, J.L.; Kennedy, S.H.; Müller, D.J.; Geraci, J. GWAS-Based Machine Learning Approach to Predict Duloxetine Response in Major Depressive Disorder. *J. Psychiatr. Res.***2018**, *99*, 62–68.
22. Zhou, J.; Theesfeld, C.L.; Yao, K.; Chen, K.M.; Wong, A.K.; Troyanskaya, O.G. Deep Learning Sequence-Based Ab Initio Prediction of Variant

Effects on Expression and Disease Risk. *Nat. Genet.***2018**, *50*, 1171–1179.
23. Storn, R.; Price, K. Differential Evolution—A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.***1997**, *11*, 341–359.
24. Price, R.; Storn, K.; Lampinen, R.M. *Differential Evolution: A Practical Approach to Global Optimization*; Springer: New York, NY, USA, 2005.
25. Huang, G.B.; Wang, D.H.; Lan, Y. Extreme Learning Machines: A Survey. *Int. J. Mach. Learn. Cybern.***2011**, *2*, 107–122.
26. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, Budapest, Hungary, 25–29 July *2004*; IEEE: Piscataway Township, NJ, USA, 2005.
27. Huang, G.B.; Zhu, Q.Y.; Siew, C.-K. Extreme Learning Machine: Theory and Applications. *Neurocomputing***2006**, *70*, 489–501.
28. Deng, W.; Zheng, Q.; Chen, L. *Regularized Extreme Learning Machine*. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 30 March–2 April 2009; IEEE: Piscataway Township, NJ, USA, 2009.
29. Joshi, G.P.; Alenezi, F.; Thirumoorthy, G.; Dutta, A.K.; You, J. Ensemble of Deep Learning-Based Multimodal Remote Sensing Image Classification Model on Unmanned Aerial Vehicle Networks. *Mathematics***2021**, *9*, 2984.
30. Gupta, U.; Gupta, D. Regularized Based Implicit Lagrangian Twin Extreme Learning Machine in Primal for Pattern Classification. *Int. J. Mach. Learn. Cybern.***2021**, *12*, 1311–1342.
31. Prakapenka, D.; Liang, Z.; Jiang, J.; Ma, L.; Da, Y. A Large-Scale Genome-Wide Association Study of Epistasis Effects of Production Traits and Daughter Pregnancy Rate in U.S. Holstein Cattle. *Genes* **2021**, *12*, 1089.
32. Gondro, C.; van der Werf, J.; Hayes, B. *Genome-Wide Association Studies and Genomic Prediction;* Methods in Molecular Biology; Humana Press: New York, NY, USA, 2013.
33. Marozzi, M. A bi-aspect nonparametric test for the two-sample location problem. *Comput. Stat. Data. Anal.***2002**, *64*, 639–648.
34. Marozzi, M. Some remarks about the number of permutations one should consider to perform a permutation test. *Statistica***2004**, *64*, 193–201.
35. Browning, B.L. PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinform.***2008**, *9*, 309.

36. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic. Acids. Res.***2000**, *28*, 27–30.
37. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein. Sci.***2019**, *28*, 1947–1951.
38. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic. Acids. Res.***2021**, *49*, D545–D551.
39. Liu, X.; Gao, C.; Li, P. A comparative analysis of support vector machines and extreme learning machines. *Neural Networks***2012**, *33*, 58–66.
40. Cheng, G.J.; Cai, L.; Pan, H.X. Comparison of extreme learning machine with support vector regression for reservoir permeability prediction. In Proceedings of the 2009 International Conference on Computational Intelligence and Security, Beijing, China, 11–14 December 2009; IEEE: Piscataway Township, NJ, USA, 2009; Volume 2, pp. 173–176.
41. Huang, G.B.; Ding, X.; Zhou, H. Optimization method based extreme learning machine for classification. *Neurocomputing* **2010**, *74*, 155–163.
42. Price, K.V.; Storn, R. Differential evolution—A simple evolution strategy for fast optimization. *Dr. Dobbs. J.***1997**, *22*, 18–24.
43. Tušar, T.; Filipi, B. Differential Evolution versus Genetic Algorithms in Multiobjective Optimization. In: *Evolutionary Multi-Criterion Optimization*, *Matsushima, Japan, 2007*; Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T., Eds.; Springer: Berlin, Heidelberg, 2007; Volume 4403.
44. Thomas, M.; Sakoda, L.C.; Hoffmeister, M.; Rosenthal, E.A.; Lee, J.K.; van Duijnhoven, F.J.B.; Platz, E.A.; Wu, A.H.; Dampier, C.H.; de la Chapelle, A. Genome-Wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *Am. J. Hum. Genet.***2020**, *107*, 432–444.
45. Yang, Y.; Lv, S.Y.; Ye, W.; Zhang, L. Apelin/APJ system and cancer. *Clin. Chim. Acta.***2016**, *457*, 112–116.
46. Picault, F.X.; Chaves-Almagro, C.; Projetti, F. Tumour co-expression of apelin and its receptor is the basis of an autocrine loop involved in the growth of colon adenocarcinomas. *Eur. J. Cancer.***2014**, *50*, 663–674.
47. Mughal, A.; O'Rourke, S.T. Vascular effects of apelin: Mechanisms and therapeutic potential. *Pharmacol. Ther.***2018**, *190*, 139–147.
48. Podgórska, M.; Diakowska, D.; Pietraszek-Gremplewicz, K.; Nienartowicz, M.; Nowak, D. Evaluation of Apelin and Apelin Receptor Level in the Primary Tumor and Serum of Colorectal Cancer Patients. *J. Clin. Med.***2019**, *8*, 1513.
49. Podgórska, M.; Pietraszek-Gremplewicz, K.; Olszanska, J.; Nowak, D. The Role of Apelin and Apelin Receptor Expression in Migration and Invasiveness of Colon Cancer Cells. *Anticancer. Res.* **2021**, *41*, 151–161.

50. Sanchez Pino, M.J.; Moreno, P.; Navarro, A. Mitochondrial dysfunction in human colorectal cancer progression. *Front. Biosci.***2007**, *12*, 1190–1199.
51. Guo, B.; Han, X.; Tkach, D.; Huang, S.G.; Zhang, D. AMPK promotes the survival of colorectal cancer stem cells. *Anim. Models Exp. Med.***2018**, *1*, 134–142.
52. Khabaz, M.N.; Abdelrahman, A.S.; Al-Maghrabi, J.A. Expression of p-AMPK in colorectal cancer revealed substantial diverse survival patterns. *Pak. J. Med. Sci.***2019**, *35*, 685–690.
53. Wu, Z.; Liu, Z.; Ge, W.; Shou, J.; You, L.; Pan, H.; Han, W. Analysis of potential genes and pathways associated with the colorectal normal mucosa-adenoma-carcinoma sequence. *Cancer. Med.***2018**, *7*, 2555–2566.
54. Yagi, T.; Kubota, E.; Koyama, H.; Tanaka, T.; Kataoka, H.; Imaeda, K.; Joh, T. Glucagon promotes colon cancer cell growth via regulating AMPK and MAPK pathways. *Oncotarget***2018**, *9*, 10650–10664.
55. Murmann, A.E.; Gao, Q.Q.; Putzbach, W.E.; Patel, M.; Bartom, E.T.; Law, C.Y.; Bridgeman, B.; Chen, S.; McMahon, K.M.; Thaxton, C.S.; et al. Small interfering RNA s based on huntingtin trinucleotide repeats are highly toxic to cancer cells. *EMBO Rep.***2018**, *19*, e45336.
56. Pechlivanis, S.; Pardini, B.; Bermejo, J.L.; Wagner, K.; Naccarati, A.; Vodickova, L.; Novotny, J.; Hemminki, K.; Vodicka, P.; Försti, A. Insulin pathway related genes and risk of colorectal cancer: INSR promoter polymorphism shows a protective effect. *Endocr.-Relat. Cancer***2007**, *14*, 733–740.
57. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832.
58. Aslam, N. Explainable Artificial Intelligence Approach for the Early Prediction of Ventilator Support and Mortality in COVID-19 Patients. *Computation***2022**, *10*, 36.
59. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable AI: The New 42? In *Machine Learning and Knowledge Extraction*; Springer: Cham, Switzerland, 2018; pp. 295–303.
60. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
61. Iwendi, C.; Bashir, A.K.; Peshkar, A.; Sujatha, R.; Chatterjee, J.M.; Pasupuleti, S.; Mishra, R.; Pillai, S.; Jo, O. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front. Public Health***2020**, *8*, 357.
62. Rezaei-Ravari, M.; Eftekhari, M.; Saberi-Movahed, F. Regularizing extreme learning machine by dual locally linear embedding manifold

learning for training multi-label neural network classifiers. *Eng. Appl. Artif. Intell.***2021**, *97*, 104062.

ARTICULO VII BIBLIOGRAFIA:

1. Ahmed Fouad Kotb*et al.* Sexual activity and the risk of Prostate Cancer: Review Article. Arch Ital UrolAndrol 2015.
2. Barnard RJ. Prostate cancer prevention by nutritional means to alleviate metabolic syndrome. Am J Clin Nutr (2007); 86s:889S-93S.
3. Blows FM et al. Subtyping of breast cancer by inmunohisto-chemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med (2010);25;7(5):e1000279.
4. Breslow A. et al. Review of Epidemiologic Studies of Alcohol and Prostate Cancer: 1971-1996. Nutr Cancer. 1998
5. Calle EE et al. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. N R Cancer (2004); 4:579-591.
6. Campos Y, Martin MA, Arenas J. *Genética molecular de las enfermedades de la cadena respiratoria mitocondrial.*Rev Neurol 2002; 35:153-8.
7. Catherine M Bulka*et al.* Arsenic in Drinking Water and Prostate Cancer in Ilinois Counties: An Ecologic Study Environ Res 2016.
8. Cosimo De Nunzio*et al.*Cigarette Smoking Is not associated with Prostate Cancer Diagnosis and agressiveness: A Cross Sectional Italian Study. Minerva UrolNefrol Dec 2018
   CurrUrol Resp. 2019
   Evaluating the association between artificial light al night exposure and Breast and Prostate Cáncer Risk in Spain (MCC Spain Study). Environ Health Pers. 2018

Exposure to pesticides and prostate cáncer: systematic review of the literatura. Rev. EnvironHealth. 2016. Review.

Factores dietéticos asociados al Cáncer de Próstata. Beneficios de la dieta Mediterránea. Actas Urol. Esp. 2012 ; 36(4): 239-245

9. Fish consumption and prostate cancer risk and mortality in a Danish Cohort Study. Europ. J. Cancer. Prev. Jul. 2018.
10. Freyre-Bernal, Sofia Isabel; Saavedra-Torres J.S; Zuñiga-Cerón L. F.*Cáncer y mitocondria. Un aspecto central para el desarrollo y cremiento tumoral.* Medicina 2017. 39(1).17-35.
11. Friedenreich CM et al. State of the epidemiological evidence on physical activity and cancer prevention. European Journal of Cancer (2010); 46:2593-2604.
12. Friedman J.R, Nunnari J. *Mitocondrial form and functions.* Nature (505) 335-343. Sutovsky P, Moreno R.D, Schatten G.
13. Gallagher EJ et al. Minireview: IGF, Insulin, and cancer. Endocrinology (2011);152:2546-2551.
14. Garcia Silva MT. *Clasificación y aspectos clínicos de las enfermedades mitocondriales.*AnEspPediatr 1996; 83: 285-290.
15. García-Saenz A. et al.
16. Gutiérrez-Fisac JL *et al*. Prevalence of general and abdominal obesity in the adult population of Spain. 2008-2010; The ENRICA Study. Obes Rev 2012 April 13(4): 388-92
17. Ludmila Prokunina-Olson, Yi-Ping Fu, Wei Tang. *Refining the Prostate Cancer genetic association within the JAZF1 gene on chromosome 7p15.2* Cancer Epidemiol Biomarkers Prev. 2010 May(5); 1349-55.
18. J. Ferris-Tortajada, U. Berbel-Torneiro, J. García Castell et al.
19. Kiefel BR, Gilson PR, Beech Pl. 2006. *Cell biology of mitochondrial dynamics. International review of cytology.* 254: 151-213.
20. Kolawole, SeiduMoshood, "Variantes genéticas asociadas con el cáncer de próstata metastásico" (2012). *Disertaciones del Centro Médico de Texas (a través de ProQuest).* AAI1515606. https://digitalcommons.library.tmc.edu/dissertations/AAI1515606
21. Li et al. A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat (2011);126:717-727.

22. MacAskill AF, Kittler J.T. *Control of mitochondrial transport and location in neurons.* Trends in cell biology. 2010. 20:102-112.
Main and interactive effects of physical activity, fitness and body mass in the prevention of cancer from the Copenhage Male Study.
23. Nuñez C, Johan Clausen, Magem Thorsten *et al.*
24. Palou A., ML Bonet, C. Picó, AM Rodríguez . Nutrigenómica y obesidad. Rev. Med. Universidad de Navarra. 2004. 36-38
25. Paulides S. Tsrigos A. Vera T. *Pruebas transcripcional del "Reverse WarburgEffect" en el estroma del tumor del cáncer de mama y metástasis. Semejanzas con estrés oxidativo, enfermedad de Alzheimer y neuroglia metabolismo de acoplamiento. Envejecimiento.* Albany N.Y. 2(4) 185-99.
26. Peters CE et al. Ocupational exposure to solar ultraviolet radiation and the risk of prostate cancer. Ocup Environ Med. 2016
27. Plathow C, Weber W.A. *Tumor cell metabolism imagin.* J. Nucl. Med. 2008 49(Supl. 2) 43-63
28. Porter C. M. et al.
29. Prostate Cancer Epidemiology. Arch Esp Urol. 2014. Review.
30. Rui Peng *et al.* Does Exposure to Asbestos Cause Prostate Cancer? A Systematic Literature Review and Meta-analisis. Medicine (Baltimore) Jan 2019
Sci Rep 2018;8-11780
31. Shui *et al.*Trichomonas Vaginaleis Infection and risk of Advanted Prostate Cancer. Prostate. May 2016.
32. Silva J. F*. et al.*
The microbiome and Prostate Cancer Risk.
The microbiome in Prostate inflammation and prostate cáncer. Prostate cancer diseases. 2018. Review.
33. *Ubiquitin tag for spemmitocondria.* Nature. 1999. 402:371.
34. Volgareva GM *et al.*ProstateCancer: Papillomaviruses as a Possible Cause. 2015.
35. Wang et al. Pathway-based approaches for analysis of genome wide association studies. Am J Hum Genet (2007); 81.doi:10.1086/522374.

36. Warburg O. *On the origin on the cancer cells.*Science. 123. 309-314 (1956).
37. Wheeler K.M. *et al.*
38. Xingwang Ye. Shilpa N. Bhupathiraju and Katherine L. Tucker. Variety in fruit and vegetable intake and cognitive function in middle and older Puerto Rican adults. British Journal of Nutrition. 2013. 109, 503-510.
39. Yuanle Deng *et al.* The Extract From Punica Granatum (Pomegranate) Peel Induces Apoptosis and Impairs Metastasis in Prostate Cancer Cells. Biomed Pharmacother. Sept 2017.