



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

## Trabajo Fin de Máster

Máster Universitario en Ingeniería y Tecnología de Sistemas Software.

Departamento de Sistemas Informáticos y Computación

# Un proceso para la identificación sistemática de variaciones genómicas: Aplicaciones a la medicina de precisión

Septiembre de 2018, Valencia

Dirigido por:

**Prof. Dr. Óscar Pastor López**

*opastor@dsic.upv.es*

Septiembre 2018

Directora Experimental:

**Ana León**

*aleon@pros.upv.es*

Septiembre 2018

Alumno:

**Simranpreet Kaur**

[simkau@inf.upv.es](mailto:simkau@inf.upv.es)



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA

Un proceso para la identificación sistemática de variaciones  
genómicas: Aplicaciones a la medicina de precisión

A process for a systematic identification of genomic variations: Applications to precision medicine

Un proces per a l'identificació sistemàtica de variacions genòmiques: Aplicacions a la medicina de precisió

El continuo descubrimiento de nueva información genómica genera un enorme volumen de datos relacionados con mutaciones genéticas que pueden ser relevantes para el diagnóstico clínico genómico de la enfermedad analizada. La identificación precisa y correcta de qué variaciones son las significativas a efectos de dicho diagnóstico es un problema de primera magnitud en el ámbito de la moderna Medicina de Precisión

Este proyecto enfrenta y propone una solución a ese problema: como determinar qué variaciones son las correctas, teniendo que seleccionarlas entre un conjunto extenso y diverso de fuentes de datos genómicos, con información muchas veces inconsistente, incompleta, presentados en formatos diversos, heterogénea y en definitiva, de complejo tratamiento tanto por el volumen de datos implicado como por la comentada heterogeneidad en la procedencia de los datos que hay que gestionar. Dicha heterogeneidad en los repositorios y la variedad de datos existentes generan conflictos durante la interpretación. El trabajo desarrollado en el proyecto unifica y sistematiza el proceso de identificación para asegurar la fiabilidad y precisión por su futura aplicación en la medicina clínica.

Para conseguirlo se utiliza la metodología SILE para la búsqueda, identificación, carga e explotación de los datos genómicos. El trabajo desarrollado ha explorado en profundidad la fase de identificación, mejorándolo de forma continua y analizando su viabilidad con dos casos prácticos que han demostrado la viabilidad del proceso de identificación. Se ha comprobado también que se cumplen con los criterios de calidad que la metodología SILE propone. El trabajo de Tesis fin de máster se ha aplicado a nivel práctico a la búsqueda e identificación de los SNPs (Single Nucleotide Polimorphism), que se refieren a Polimorfismos de Nucleótidos que están asociados a la enfermedad de Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda.

El proceso sistemático de identificación de variaciones genómicas es un paso adelante concreto y relevante en el ámbito del desarrollo del cuerpo de conocimiento asociado al diseño y gestión de Sistemas de Información Genómicos, líneas de I+D muy activa en el Centro de I+D en Métodos de Producción de Software de la Universidad Politécnica de Valencia, y conecta de forma precisa Ingeniería de Sistemas de Información, Ciencias de Datos Genómicos y aplicaciones a la Medicina de Precisión.

Palabras claves: SNP, variaciones genómicas, medicina de precisión, bases de datos genómicas, SILE

## *Abstract*

The continuous discovery of new genomic information generates a huge volume of data related to genetic mutations that may be relevant to the clinical genomic diagnosis of diseases. The precise and correct identification of significant variations for the diagnosis of indicated purpose is a problem of first magnitude in the field of modern Precision Medicine

This project confronts and proposes a solution to this problem: how to determine which variations are the right ones, selecting them among an extensive and diverse set of genomic data sources, with information that is often inconsistent, incomplete, presented in diverse formats, heterogeneous and definitive, complex treatment both for the volume of data involved and for the aforementioned heterogeneity in the origin of the data that must be managed. This heterogeneity in the repositories and the variety of existing data generates conflicts during the interpretation. The work developed in the project unifies and systematizes the identification process to ensure reliability and accuracy for its future application in clinical medicine.

To achieve this, the SILE methodology is used to search, identify, load and exploit genomic data. The work developed has explored in depth the identification phase, improving it continuously and analyzing its viability with two practical cases that have demonstrated the viability of the identification process. It has also been verified that the quality criteria that the SILE methodology proposes are met. This Master's Thesis work has been applied at a practical level to search and identify SNPs (Single Nucleotide Polymorphism), which refer to Nucleotide Polymorphisms that are associated with the disease of Type 2 Diabetes Mellitus and Deep Vein Thrombosis.

The systematic process of identifying relevant genomic variations is a concrete and relevant step forward in the development of the body of knowledge associated with the design and management of Genomic Information Systems, very active R & D field in the R & D Center in Software Production Methods of the Polytechnic University of Valencia, and precisely connects Information Systems Engineering, Genomic Data Sciences and applications to Precision Medicine.

Keywords: SNP, genomic variations, Precision Medicine, genomic data base, SILE

## Tabla de Contenido

Capítulo 1. Introducción .....	10
1.1. Motivación .....	10
1.2. Objetivos del trabajo.....	11
1.3. Estructura del documento.....	12
Capítulo 2. Estado del arte.....	14
2.1. Introducción .....	14
2.2. Proyectos importantes.....	16
2.2.1. Proyecto genoma humano (PGH).....	16
2.2.2. Proyecto internacional HapMap.....	17
2.2.3. Proyecto ENCODE .....	18
2.2.4. Proyecto 1000 genomas.....	19
2.3. Conceptos importantes.....	21
2.3.1. ADN.....	22
2.3.2. Genoma Humano .....	23
2.3.3. Gen.....	24
2.3.4. Proteínas codificadas a partir de adn .....	25
2.3.5. Alelos .....	26
2.3.6. Mutaciones.....	27
2.3.7. Haplotipo.....	28
2.3.8. Fenotipo.....	28
2.3.9. Secuenciación .....	29
2.3.10. Medicina de Precisión .....	29
Capítulo 3. Enfermedades a tratar .....	30
3.1. Diabetes mellitus tipo 2 .....	30
3.2. Trombosis venosa profunda .....	32
Capítulo 4. Repositorios tratados .....	35
4.1. Ensembl .....	36
4.2. ClinVar .....	37
4.3. SNPedia .....	39
4.4. GWAS Catalog.....	40
4.5. Estructura de datos utilizada.....	41
Capítulo 5. Metodología SILE .....	44
5.1. Introducción .....	44
5.2. Search.....	44

5.3.	<i>Identification</i> .....	45
5.3.1.	Criterios de calidad.....	45
5.3.1.1.	Significado clínico .....	45
5.3.1.2.	Número de publicaciones asociadas.....	46
5.3.1.3.	Criterios de calidad .....	46
5.3.2.	Inconsistencias en los datos .....	48
5.3.2.1.	Variaciones de enfermedades relacionadas.....	48
5.3.2.2.	Afectado por genes de proximidad.....	50
5.3.2.3.	Estudio de efecto SNP por SNP .....	52
5.3.2.4.	Falta de Sujetos de estudio .....	53
5.3.2.5.	Falta de valores Estadísticos pertenecientes a SNPedia.....	53
5.3.2.6.	Citaciones .....	54
5.3.2.7.	Población y fases de estudios.....	54
5.3.3.	Criterios de calidad con los datos completos .....	55
5.4.	Preparación anterior a la carga .....	55
Capítulo 6. Conclusiones y trabajos futuros.....		57
Bibliografía.....		59
Anexo 1. Parámetros finales para la carga de datos.....		63
Anexo 2. Preparación de datos para la Carga.....		65
Anexo 3. Resumen Proceso Identificación .....		67

## Índice de Ilustraciones

ILUSTRACIÓN 1. REPOSITARIOS CONSULTADOS: ENSEMBL, CLINVAR, SNPEDIA Y GWAS CATALOG ..	11
ILUSTRACIÓN 2. TERAPIA GÉNICA [54] .....	15
ILUSTRACIÓN 3. DIAGNÓSTICO PREIMPLANTACIONAL [55] .....	15
ILUSTRACIÓN 4. TÉCNICA DE SECUENCIACIÓN AUTOMÁTICA DEL GENOMA HUMANO [56].....	17
ILUSTRACIÓN 5. EJEMPLO DE RESULTADOS PERTENECIENTES A ESTUDIO POBLACIONAL HAPMAP [57].....	18
ILUSTRACIÓN 6. EJEMPLO ELEMENTOS MAPEADOS Y ESTUDIADOS EN EL PROYECTO ENCODE [46] .....	19
ILUSTRACIÓN 7. POBLACIÓN PARTICIPANTE EN EL PROYECTO 1000 GENOMA [29] .....	19
ILUSTRACIÓN 8. MEDICINA DE PRECISIÓN [58] .....	21
ILUSTRACIÓN 9. ADN EN CÉLULA EUCARIOTA [59].....	22
ILUSTRACIÓN 10. GENOMA HUMANO CON LOS 23 PARES DE CROMOSOMAS MASCULINO Y FEMENINO [60]	23
ILUSTRACIÓN 11. GEN COMPUESTO POR EXONES E INTRONES [61].....	24
ILUSTRACIÓN 12. DE ADN A PROTEÍNAS: TRANSCRIPCIÓN Y TRADUCCIÓN [62].....	25
ILUSTRACIÓN 13. ALELOS EN CROMOSOMAS DIPLOIDES [63] .....	26
ILUSTRACIÓN 14. MUTACIONES GÉNICAS [64].....	27
ILUSTRACIÓN 15. HAPLOTIPO EN UNA REGIÓN DEL CROMOSOMA [33].....	28
ILUSTRACIÓN 16. GENOTIPO Y SU MANIFESTACIÓN COMO FENOTIPO [65] .....	29
ILUSTRACIÓN 17. SÍNTOMAS DE DIABETES MELLITUS TIPO 2 [66].....	31
ILUSTRACIÓN 18. MAPAMUNDI CON LA ESTADÍSTICA DE CASOS DE DIABETES MELLITUS TIPO 2 [15].....	32
ILUSTRACIÓN 19. TROMBOSIS VENOS PROFUNDA [67].....	33
ILUSTRACIÓN 20. TROMBOSIS VENOSA PROFUNDA EN EL MUNDO [68] .....	34
ILUSTRACIÓN 21. REPOSITORIO ENSEMBL [69] .....	36
ILUSTRACIÓN 22. REPOSITORIO ENSEMBL REGIONES DE INTERÉS EN LA INTERFAZ WEB [70].....	36
ILUSTRACIÓN 23. REPOSITORIO CLINVAR [71] .....	37
ILUSTRACIÓN 24. LAS APORTACIONES MUNDIALES A LA BASE DE DATOS DE CLINVAR [72].....	38
ILUSTRACIÓN 25. REPOSITORIO SNPEDIA [73] .....	39
ILUSTRACIÓN 26. REPOSITORIO GWAS CATALOG [74].....	40
ILUSTRACIÓN 27. ESTADÍSTICA GWAS CATALOG 2013 [22] .....	41
ILUSTRACIÓN 28. VARIACIONES DE ENFERMEDADES RELACIONADAS ENSEMBL [75].....	48
ILUSTRACIÓN 29. VARIACIONES DE ENFERMEDADES RELACIONADAS: RESUMEN ARTÍCULO [27] .....	49
ILUSTRACIÓN 30. VARIACIONES DE ENFERMEDADES RELACIONADAS EN GWAS CATALOG [76].....	49
ILUSTRACIÓN 31. ARTÍCULO DE ENFERMEDADE RELACIONADA [77] .....	50
ILUSTRACIÓN 32. SNP LOCALIZADO FUERA DEL GEN EN ENSEMBL [78].....	50
ILUSTRACIÓN 33. SNP LOCALIZADO EN DBSNP [79] .....	51
ILUSTRACIÓN 34. SNP LOCALIZADO FUERA DEL GEN [80] .....	51
ILUSTRACIÓN 35. ESTUDIO VARIACIÓN CON INTERACCIÓN GEN POR GEN. [81] .....	52
ILUSTRACIÓN 36. EVOLUCIÓN DE ESTUDIOS Y ARTÍCULO EN GWAS [28] .....	52
ILUSTRACIÓN 37. FALTA SUJETOS DE ESTUDIO [82] .....	53
ILUSTRACIÓN 38. FALTA DE VALORES SNPEDIA [83] .....	53
ILUSTRACIÓN 39. POBLACIÓN Y SUBPOBLACIÓN [84].....	54
ILUSTRACIÓN 40. ENSEMBL DISTINTOS P-VALUES [85].....	55
ILUSTRACIÓN 41. ALELO DE REFERENCIA Y ALELO ALTERNATIVO EN ENSEMBL .....	65
ILUSTRACIÓN 42. FRECUENCIA DE REFERENCIA, FRECUENCIA DEL GEN ALTERNATIVO, FRECUENCIA GENOTIPO 1, FRECUENCIA GENOTIPO 2 Y FRECUENCIA GENOTIPO 3 EN ENSEMBL.....	65
ILUSTRACIÓN 43. NOMBRE DE LA VARIANTE EN ENSEMBL .....	66
ILUSTRACIÓN 44. REFERENCIAS DEL AUTORES EN ENSEMBL .....	66



## *Índice de tablas*

TABLA 1. ARCHIVOS PRESENTES EN 2018 EN LA BASE DE DATOS CLINVAR [41].....	38
TABLA 2. ARCHIVOS PRESENTES EN 2018 EN LA BASE DE DATOS SNPEDIA [21] .....	39
TABLA 3. PARÁMETROS DE SALIDA DESDE ENSEMBL, CLINVAR, SNPEDIA Y GWAS .....	43
TABLA 4. RESULTADOS BÚSQUEDA DIABETES MELLITUS TIPO 2 Y TROMBOSIS VENOSA PROFUNDA .....	44
TABLA 5. RESULTADOS BÚSQUEDA DIABETES MELLITUS TIPO 2 Y TROMBOSIS VENOSA PROFUNDA (*DESCARTANDO LAS FILAS CON VALORES VACIOS) .....	45
TABLA 6. RESULTADOS FINALES.....	56
TABLA 7. PARÁMETROS FINALES PARA LA CARGA DE DATOS.....	64
TABLA 8. PROCESO IDENTIFICACIÓN DIABETES MELLITUS TIPO 2 .....	67
TABLA 9. PROCESO IDENTIFICACIÓN TROMBOSIS VENOSA PROFUNDA.....	67

# CAPÍTULO 1. INTRODUCCIÓN

## 1.1. MOTIVACIÓN

Diagnosticar ciertas variaciones en el gen que pueden llegar a provocar mutaciones benignas o malignas es lo que inició y mueve la bioinformática. Tener variación genética benigna es una ventaja para quien lo padezca pero en el caso de tener mutación genética maligna puede llegar a provocar situaciones de gravedad para la salud. Mediante la bioinformática podemos identificar a tiempo estas mutaciones y llegar a imponer medidas preventivas en nuestra vida diaria para evitar la manifestación de esa enfermedad o ralentizar la evolución de dicha enfermedad.

Con los avances continuos en la medicina, tenemos mayores posibilidades de conocer en detalle la salud individual de cada persona mediante nuevos métodos como los estudios genéticos. La dimensión molecular de los diagnósticos genéticos es capaz de detectar pequeñas variaciones individuales (polimorfismos) en el material genético que codifica las proteínas que constituyen la base para todos los procesos biológicos humanos.

Hoy en día es normal secuenciar fragmentos de ADN o el genoma humano en paralelo (“Next-Generation Sequencing” de las siglas de Inglés NGS) e identificar variantes de nucleótidos para determinar predisposición a enfermedades genéticas. Pero esta identificación se realiza en base a las variaciones ya identificadas y sus posibles asignaciones con enfermedades. Estas pueden ser consultadas en algunas de las bases de datos genómicas o en todas.

Lo primero, consultar las variaciones ya identificadas en una base de datos si es muy plausible ya que se trata de conocer el modelo conceptual de esa base de datos y las posibles formas de extracción de datos. Pero lo segundo, consultar variaciones identificadas en todas las bases de datos produce un caos ya que las bases de datos difieren en sus modelos conceptuales y las formas de extracción de los datos, es decir, varían en la disposición, estructura e importancia de los datos ofrecidos ya que no existe un Modelo Conceptual del Genoma Humano general universal.

Esta segunda idea es tan tentadora porque ofrece disponibilidad e información verificada por varias bases de datos. Pero esto supone trabajo intensivo de expertos de los dos campos, tanto de informática como medicina genómica. Aun así, la identificación y recopilación de variaciones de forma eficiente y fiable es muy importante para poder aplicarlo con seguridad a los pacientes y dar un paso hacia la medicina personalizada.

La medicina personalizada o de precisión, cuyo objetivo es mejorar la eficacia para cada paciente y eliminar o disminuir los efectos secundarios de otras terapias más generalistas. Estudiando nuestro genoma se pueden llegar a localizar variaciones genómicas específicas pertenecientes a enfermedades y de esta forma poder realizar diagnósticos concretos en enfermedades. Y estamos bastante cerca de esta medicina personalizada ya que en el último medio siglo se ha dado un paso importante en la secuenciación genómica poblando masivamente las Bases de Datos Genómicas con los resultados de las identificaciones de variaciones genómicas.



ILUSTRACIÓN 1. REPOSITARIOS CONSULTADOS: ENSEMBL, CLINVAR, SNPEDIA Y GWAS CATALOG

Para ofrecer medicina personalizada debemos optar por la segunda opción de la consulta unificada a las bases de datos existentes. El gran reto aquí es la unificación así como resolver los conflictos estructurales entre los datos. En este TFM se van a tratar 4 repositorios (Ilustración 1) genómicos indicados en la Ilustración 1, en busca de variaciones que son Polimorfismos de Nucleótidos Simples, SNPs (Single Nucleotide Polimorphism)<sup>1</sup> relacionados con las enfermedades de Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda y realizar un primer paso en el proceso de unificación. Para ello se sigue el siguiente proceso aplicando el paso de 'S' búsqueda e 'I' identificación de la metodología SILE (*Search-Identification-Load-Exploitation*). Se van a tratar solo SNPs porque estas representan el 90% de las mutaciones genéticas existentes.

## 1.2. OBJETIVOS DEL TRABAJO

Para entender e interpretar correcta y completamente el ADN la genómica pone su enfoque en las bases moleculares ya que se debe empezar por la base para poder entenderlo por completo y de esta forma tratarlo de la mejor forma posible. Todos estos datos moleculares se almacenan en las bases de datos con el objetivo de:

- Conocer mejor la variabilidad genética.
- Disponer de información suficiente para reaccionar a tiempo ante enfermedades genéticas.
- Desarrollo de Medicina de Precisión.
- Identificación de patrones genéticos en los pacientes.
- Tratamiento personalizado con medicina preventiva y poder estipular la edad crítica, trastorno y la gravedad en pacientes.
- Mediante medicina personalizada poder afrontar complicaciones relacionadas con cada una de las enfermedades genómicas.

A pesar de los avances tecnológicos el reto principal está en la variedad de los repositorios y en la complejidad de gestión de datos y que la forma dispersa y diversa que tienen las bases de

---

<sup>1</sup> Aunque no siempre se trata de polimorfismos de un solo nucleótido en ocasiones pueden estar más de un nucleótido involucrado.

datos genera la problemática de interoperabilidad y la utilización paralela de varias bases de datos existentes.

En este TFM se seguirá la metodología SILE desarrollada en el Centro de Investigación PROS de la Universidad Politécnica de Valencia. Está compuesta por los siguientes 4 pasos:

1. *Search*: Búsqueda de fuentes de datos de variaciones genéticas, en este TFM de un solo nucleótido SNPs (del inglés “*Single Nucleotide Polymorphisms*”) asociados directamente a Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda.
2. *Identification*: Identificación de las variaciones genéticas mediante aplicación de filtros y validación de datos con criterios científicos.
3. *Load*: Carga selectiva de los datos para consultas futuras.
4. *Exploitation*: Explotación de los datos para la obtención de información y resultados.

En este TFM se va centrar más en el paso 2 ya que se obtiene una cantidad masiva de datos solo de SNPs pero pocas de ellas con aportación significativa en un diagnóstico genómico. Se sigue un proceso de filtrado y estudio de los datos para validarlos. Concluyendo, los objetivos de este Trabajo Fin de Máster son:

- Emplear la metodología SILE para la búsqueda e identificación de datos para que queden en estado de poder cargar a un repositorio unificado.
- Estudiar las bases de datos genómicas para entender su modelado de datos.
- Búsqueda de variaciones genéticas que solo afectan a un nucleótido (SNP).
- Identificar los SNPs válidos para Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda
- Procesar los SNPs siguiendo criterios de calidad para una mejor gestión.

### 1.3. ESTRUCTURA DEL DOCUMENTO

Con la información existente en las bases de datos genómicas sobre Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda se realiza una recopilación de las variaciones genómicas que afectan a un solo nucleótido, SNP. Se recogen variaciones de ClinVar, Ensembl, GWAS Catalog y SNPedia que utilizan formatos distintos para las variaciones genéticas. El TFM consiste en entender y unificar estos formatos así como plasmar toda la información necesaria de las variaciones genómicas encontradas en estos cuatro repositorios.

Se aplicaron criterios de calidad para considerar mutaciones relevantes y se realizó un barrido completo de los datos y variaciones. Con estos pasos nos quedamos con variaciones significantes para cada enfermedad y para acabar se añadieron datos complementarios para cada variación que cumplieran con el modelo proporcionado por la directora experimental.

En resumidas cuentas en este Trabajo Fin de Máster se seguirá la siguiente estructura:

- En el capítulo 2, se introduce el estado del arte para una primera toma de contacto con el campo de medicina genómica. Y se describen los conceptos que fueron apareciendo durante el desarrollo. A su vez se realizan explicaciones breves de términos y procesos genómicos para una mejor comprensión de los datos utilizados.
- En el capítulo 3, se realiza una presentación de las dos enfermedades a tratar Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda las dos muy presentes en la península ibérica.
- En el capítulo 4, se realiza una presentación de los repositorios existentes, así como se describen las bases de datos y los repositorios consultadas durante todo el proceso de búsqueda e identificación de variaciones fenotípicas.
- En el capítulo 5, se desarrollan los dos primeros pasos de la metodología SILE para Diabetes Mellitus Tipo 2 (DMT2) y Trombosis Venosa Profunda (TVP). Este capítulo se desglosa en la búsqueda de variaciones fenotípicas en Ensembl, ClinVar, GWAS Catalog y SNPedia. Posteriormente se realiza la identificación y procesamiento exhaustivo de los datos genómicos y se muestran las inconsistencias encontradas durante el proceso de identificación de los datos. Indicando los pasos de procesamiento aplicados a los casos y las decisiones tomadas junto con sus repercusiones.
- En el capítulo 6, se redactan las conclusiones sobre el trabajo realizado y finalmente las propuestas de trabajos futuros.

## CAPÍTULO 2. ESTADO DEL ARTE

### 2.1. INTRODUCCIÓN

El genoma humano es la secuenciación de los 23 pares de cromosomas que se encuentran en nuestras células y realizan las funciones que son la base de la vida. Desde la secuenciación completa del genoma llevada a cabo por el Proyecto Genoma humano [2] en 2016 hasta ahora estamos en proceso de poder decodificar por completo el comportamiento de todos los genes en nuestro ADN. En el Proyecto Genoma Humano se identificaron de 26 mil hasta 30 mil genes humanos. Estos genes son de gran importancia ya que son la base de la medicina genética y biomédica. Los datos de genes mapeados en el Proyecto de Genoma Humano son de acceso web público disponible.

Se está trabajando intensivamente en la determinación de las funciones sobre los genes mapeados así como tratar las mutaciones en esos genes. Con el mapeo correcto se puede llegar a conocer todos los fenotipos asociados, es decir, variaciones genéticas que provocan enfermedad o enfermedades genómicas. El mapeo e identificación de los fenotipos es lo que haría posible la identificación de una enfermedad en un diagnóstico genético.

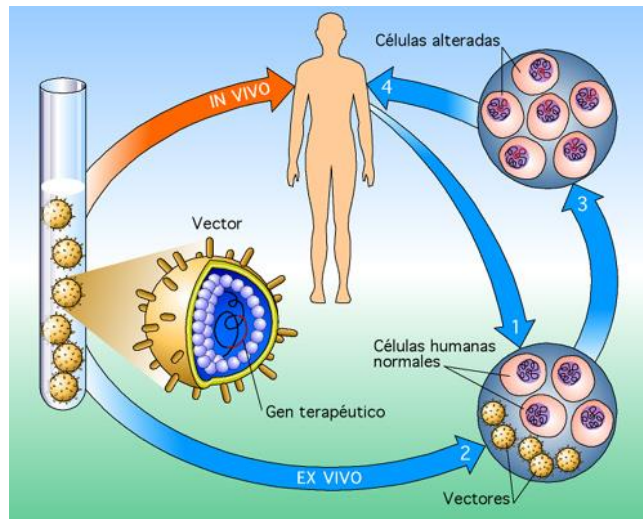
Con el conocimiento de las bases moleculares repartidas por el mundo se ha podido realizar diagnósticos adecuados de muchas enfermedades como es el caso de Enfermedad de Gaucher, Enfermedad de Alzheimer, Enfermedad de Huntington y Síndrome de Marfan [2].

Se pueden realizar diagnósticos pre sintomático en caso de tener predisposición genética a una enfermedad y llegar a tomar medidas preventivas para el control de otros factores que agudicen o aceleren la aparición de la enfermedad. Y por último el diagnóstico prenatal ya es un caso práctico aplicado en muchos de los hospitales en el cual también se han llegado a diagnosticar enfermedades genéticas. Pero para ello se deben tener mapeados e identificadas variaciones fenotípicas asociadas a la enfermedad en bases de datos de fácil manejo y disponibilidad.

Otra rama importante es la terapia que se sigue después del diagnóstico. En casos positivos de diagnósticos se deben tomar medidas relacionadas a la prevención y disminución respecto a la gravedad de la enfermedad. Existen distintas opciones de tratamiento dependiendo del caso [2]:

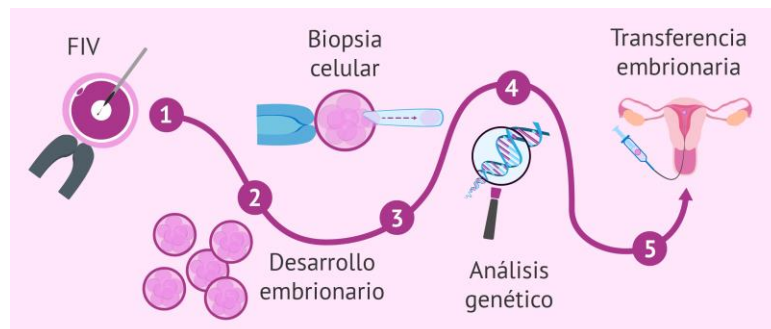
- Terapia Génica: una técnica en desarrollo que consiste en la inserción de genes modificados en el metabolismo del paciente pero esta terapia solo se puede aplicar en ciertas condiciones y casos. Por ejemplo la regulación del gen sea

precisa y conocida. Además, se puede llegar a realizar terapia génica somática<sup>2</sup> o germinal<sup>3</sup>. La Ilustración 2 muestra la metodología utilizada.



**ILUSTRACIÓN 2. TERAPIA GÉNICA [54]**

- Terapia Farmacológica: consiste en neutralizar las alteraciones mediante medicamentos que eliminen o minimicen los efectos secundarios de la enfermedad. Para ello se crea un perfil genético del paciente y esto se ha hecho posible por el Proyecto Genoma Humano y entre otros proyectos de mapeo e identificación de variaciones genéticas.
- Medicina Preventiva: realiza estudios de sujetos poblacionales para determinar los genes que provocan enfermedades determinadas y poder predecir qué medidas las van a acelerar o ralentizar. Pero esta se ve frenado por conflictos en Aspectos Éticos.
- Diagnóstico genético preimplantacional (DGPI): como el propio nombre indica se trata de realizar un diagnóstico genético a los embriones durante la fecundación in vitro. Se puede ver en resumen el proceso en la Ilustración 3.



**ILUSTRACIÓN 3. DIAGNÓSTICO PREIMPLANTACIONAL [55]**

<sup>2</sup> Terapia génica somática: Modificaciones genéticas que solo tienen lugar en dicho paciente.

<sup>3</sup> Terapia génica germinal: Modificaciones genéticas que son hereditarias y cabe la posibilidad de transmitir las a los descendientes.

## 2.2. PROYECTOS IMPORTANTES

Se han llevado a cabo numerosos proyectos relacionados con el mapeo e identificación de mutaciones pero se estima que hay alrededor de 1 a 2 millones de SNPs [2] en el genoma humano. Estos son objeto de muchos estudios y proyectos ya que podrían estar asociados a enfermedades o provocar un mayor riesgo a enfermedades, por consiguiente a estas mutaciones conocidas como SNP se les puede llegar a utilizar como marcadores para localizar genes responsables en las enfermedades. Vamos a introducir algunos de los proyectos realizados internacionalmente que enriquecieron a gran escala los repositorios genómicos:

### 2.2.1. PROYECTO GENOMA HUMANO (PGH)

Ya se ha hablado del Proyecto Genoma Humano en el cual se invirtieron hasta 3000 millones de dólares, fundado en 1990 en el Departamento de Energía y Ciencias Tránsicas y los Institutos Nacionales de la Salud de los Estados Unidos. Con los objetivos principales de secuenciar y mapear el genoma humano para su posible uso en investigaciones y estudios junto con su posterior aplicación en diagnósticos presintomáticos y prenatales.

Cabe destacar que la información fue publicada y es de acceso libre [2]. Además se realizaron supervisiones sobre temas éticos, legales y sociales para la libre utilización de los datos. Gracias a este proyecto se conoce que unos 4 mil trastornos genéticos se deben a genes mutantes, algunos muy comunes, como la sickleemia, la talasemia y la fibrosis quística, en los que se produce la alteración de un solo gen [45]. Lo que también remarcó mucho este proyecto fue la técnica de secuenciación utilizada y desarrollada por Frederick Sanger para este proyecto resumida en la Ilustración 4.



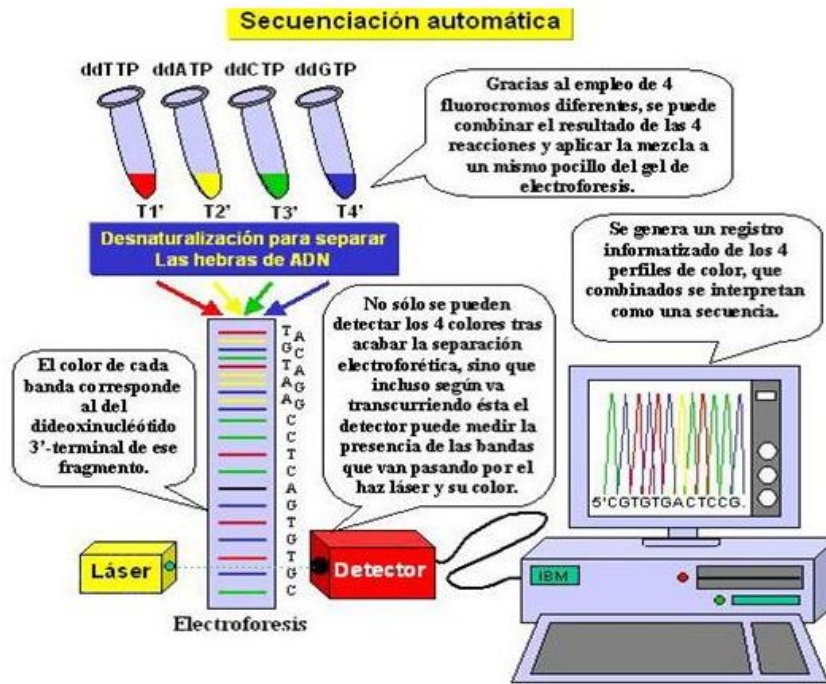


ILUSTRACIÓN 4. TÉCNICA DE SECUENCIACIÓN AUTOMÁTICA DEL GENOMA HUMANO [56]

### 2.2.2. PROYECTO INTERNACIONAL HAPMAP

Otro de los proyectos importantes ha sido el Proyecto internacional HapMap para desarrollar un mapa de haplotipos<sup>4</sup> del genoma humano y estudiar las similitudes ambientales que afectan a enfermedades genómicas mediante la catalogación regional de similitudes y diferencias genéticas poblaciones. También se asociaron SNPs (variaciones de nucleótidos simples) con enfermedades y diferentes perfiles de respuestas ante los fármacos [30].

Los datos son de acceso libre en la plataforma de NCBI (Centro Nacional para la Información Biotecnológica, proviene de la traducción de inglés '*National Center for Biotechnology Information*') y también se utilizan en repositorios web como Ensembl para las frecuencias genotípicas y frecuencias de alelos pertenecientes a distintos países como se puede observar en la Ilustración 5.

<sup>4</sup> Definido en el apartado 2.3

## HapMap Project (12) ▢








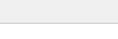
Population	Allele: frequency (count)	Genotype: frequency (count)	ssID	Submitter
<a href="#">HapMap-HCB</a>	 C: 0.384 (33) T: 0.616 (53)	C C: 0.140 (6) T T: 0.372 (16)	C T: 0.488 (21)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-ASW</a>	 C: 0.378 (37) T: 0.622 (61)	C C: 0.122 (6) T T: 0.367 (18)	C T: 0.510 (25)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-CHB</a>	 C: 0.585 (48) T: 0.415 (34)	C C: 0.268 (11) T T: 0.098 (4)	C T: 0.634 (26)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-CHD</a>	 C: 0.488 (83) T: 0.512 (87)	C C: 0.235 (20) T T: 0.259 (22)	C T: 0.506 (43)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-GIH</a>	 C: 0.523 (92) T: 0.477 (84)	C C: 0.318 (28) T T: 0.273 (24)	C T: 0.409 (36)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-LWK</a>	 C: 0.439 (79) T: 0.561 (101)	C C: 0.189 (17) T T: 0.311 (28)	C T: 0.500 (45)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-MEX</a>	 C: 0.680 (68) T: 0.320 (32)	C C: 0.460 (23) T T: 0.100 (5)	C T: 0.440 (22)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-MKK</a>	 C: 0.381 (109) T: 0.619 (177)	C C: 0.126 (18) T T: 0.364 (52)	C T: 0.510 (73)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HAPMAP-TSI</a>	 C: 0.761 (134) T: 0.239 (42)	C C: 0.580 (51) T T: 0.057 (5)	C T: 0.364 (32)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HapMap-CEU</a>	 C: 0.704 (159) T: 0.296 (67)	C C: 0.496 (56) T T: 0.088 (10)	C T: 0.416 (47)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HapMap-JPT</a>	 C: 0.360 (62) T: 0.640 (110)	C C: 0.151 (13) T T: 0.430 (37)	C T: 0.419 (36)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>
<a href="#">HapMap-YRI</a>	 C: 0.296 (67) T: 0.704 (159)	C C: 0.071 (8) T T: 0.478 (54)	C T: 0.451 (51)	<a href="#">ss44741203</a> <a href="#">CSHL-HAPMAP</a>

ILUSTRACIÓN 5. EJEMPLO DE RESULTADOS PERTENECIENTES A ESTUDIO POBLACIONAL HAPMAP [57]

### 2.2.3. PROYECTO ENCODE

ENCODE es Enciclopedia de los elementos de ADN (de las siglas de inglés “*Encyclopedia of DNA Elements*”). Es un proyecto de investigación pública lanzado por el Instituto Nacional de Investigación del Genoma Humano de Estados Unidos (NHGRI) en septiembre de 2003 y se encontraba en su cuarta fase en 2017 [6]. Destinado a ser una continuación del Proyecto Genoma Humano. Que cataloga la ubicación de los genes e identifica los elementos funcionales en el genoma humano que no son considerados de utilidad o ‘basura’<sup>5</sup> con la utilización de métodos y tecnologías variadas.

Además, en este proyecto se intenta comprender los elementos que intervienen en los procesos relacionados con el genoma humano como es el ARN (Ácido Ribonucleico), proteínas y los elementos reguladores [47]. Como se puede ver en la Ilustración 6 que resume cómo el código genético localizado en un cromosoma empieza a interactuar con otros elementos funcionales. [46]

<sup>5</sup> Sólo 1.5% del genoma humano codifica proteínas del resto 98.5% no conoce la utilidad o se considera ADN ‘basura’ o repetitivo [6]

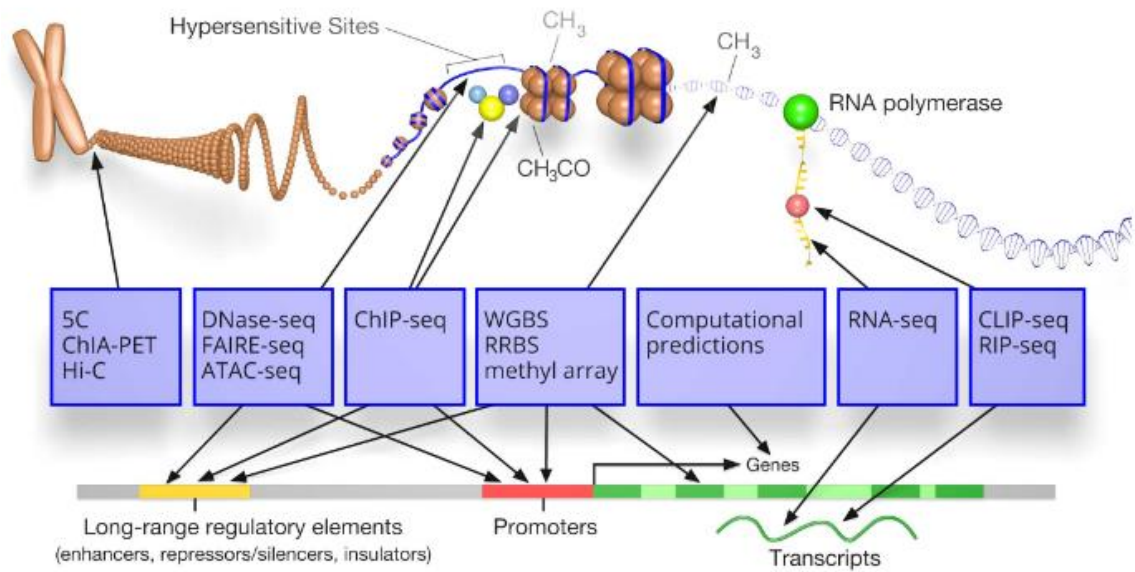


ILUSTRACIÓN 6. EJEMPLO ELEMENTOS MAPEADOS Y ESTUDIADOS EN EL PROYECTO ENCODE [46]

#### 2.2.4. PROYECTO 1000 GENOMAS

Realizó la secuenciación genómica de 2.504 individuos de 26 poblaciones distintas (se pueden apreciar en el mapa de la Ilustración 1), describe un catálogo de 84,7 millones de variantes que amplía en un 40 % el número de variantes conocidas del genoma humano según un artículo en la revista *Nature* [7]. Los resultados de las frecuencias de alelos y genotipos generados durante el proyecto están disponibles en distintos repositorios genómicos para su uso, como es el caso de Ensembl.



ILUSTRACIÓN 7. POBLACIÓN PARTICIPANTE EN EL PROYECTO 1000 GENOMA [29]

Todos estos datos sobre el genoma humano, las mutaciones genéticas y la codificación de proteínas son de acceso libre. Se debe reconocer que hay una inmensidad de información en una variedad de repositorios. Pero la cuestión aquí no es la cantidad de información a tratar sino la estructura y el formato. Al no existir un estándar para la representación de las mutaciones estas están representadas de forma dispersa y variada en los distintos repositorios.

Los datos recopilados en estos proyectos pueden ser de gran relevancia para determinar las influencias medioambientales que contribuyen a las variaciones genéticas y además la predisposición a infecciones o enfermedades y la eficacia de los medicamentos llevándonos de esta forma un paso más cerca a la Medicina de Precisión.

Aunque existen distintos tipos de repositorios, entre los más reconocidos están:

- Repositorios de secuencias de nucleótidos :
  - NCBI
  - EMBL
  - DDBJ
  - GenBank
  
- Repositorios de Enfermedades genéticas humanas :
  - OMIM
  
- Repositorios de genomas de organismos concretos :
  - Flybase
  - SGD
  - ENSEMBL
  - UCSC
  
- Repositorios de proteínas :
  - Uniprot
  - PDB
  - SCOP
  
- Repositorios de bibliografía:
  - GWAS Catalog
  - PubMed
  
- Repositorios de Polimorfismos :
  - SNPedia
  - dbSNP

Los avances tecnológicos y la intensa dedicación en este campo han generado muchos datos pero aun estamos empezando a descifrar el papel de los genes y mutaciones. Pero como se indicó anteriormente, esta información debe seguir una estructura, un estándar, tener un modelo conceptual que lo represente de forma abstracta para un fácil y ágil acceso junto con su utilización.

El Modelo Conceptual del Genoma Humano (MCGH) [4] propuesto por J. Reyes Román representa de forma sencilla los elementos genómicos para un mayor entendimiento del dominio. Esto permite modelar Sistemas de Información Genómicos eficientes para una mejor gestión de la información existente. Ya que el objetivo principal es comprender toda la información presente en el genoma humano para su posterior utilización correcta.

Un mejor entendimiento del genoma humano da el paso hacia la Medicina de Precisión. Esta práctica propone tratar a cada paciente de forma distinta teniendo en cuenta su variabilidad genética, el estilo de vida que lleva y el medioambiente habitado, como se puede ver las bases de medicina de precisión en la Ilustración 8. La genética puede determinar la predisposición a algunas enfermedades y de misma forma hay genes que responden positivamente ante un tipo de tratamiento. Esto crea mucha esperanza para las enfermedades raras y sugiere emplear tratamientos específicos.

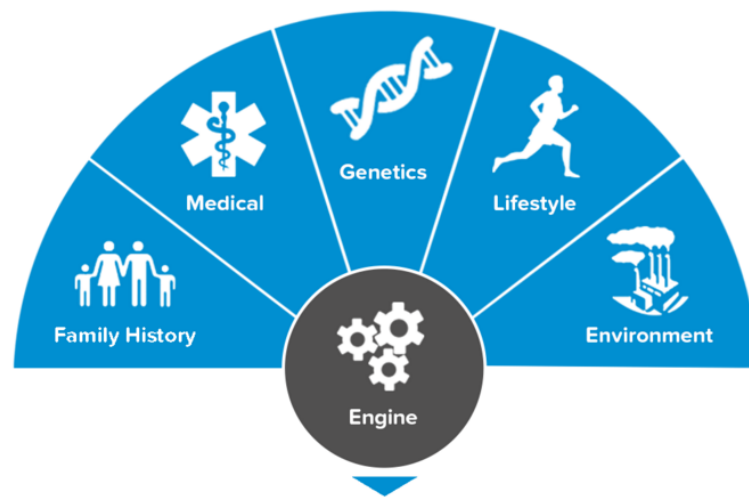


ILUSTRACIÓN 8. MEDICINA DE PRECISIÓN [58]

### 2.3. CONCEPTOS IMPORTANTES

Para poder desarrollar este trabajo se ha tenido que familiarizarse con aspectos generales de la biología genómico. Para ello es importante tener claros conceptos como es el ADN, gen, proteínas codificadas a partir del ADN, genoma humano, alelo, mutaciones, haplotipo, fenotipo, secuenciación y medicina de precisión.

### 2.3.1. ADN

El ácido desoxirribonucleico (ADN), como el propio nombre indica es un ácido nucléico que contiene la información en forma de código genético para realizar el desarrollo y funcionamientos vitales. Todos los organismos vivos almacenan el ADN en el núcleo y/o mitocondrias de cada célula eucariota. Es hereditario y su tarea principal es la creación de componentes de otras células como las proteínas y el ARN entre otros.

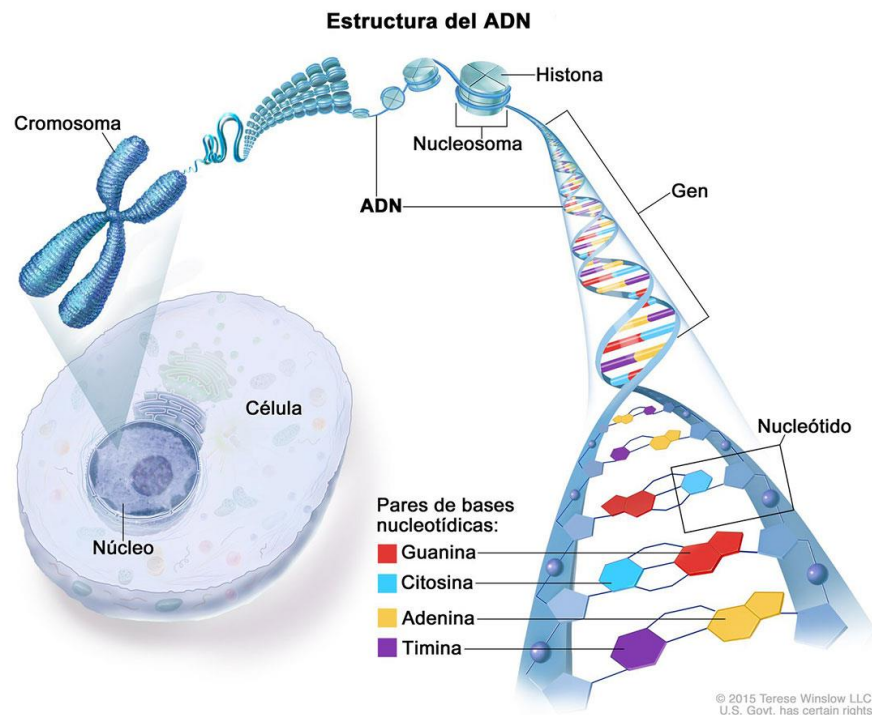


ILUSTRACIÓN 9. ADN EN CÉLULA EUCARIOTA [59]

El código genético está formado por cadena de secuencias de nucleótidos, hay 4 nucleótidos que difieren en sus bases nitrogenadas (A -> Adenina, T -> Timina, C -> Citosina y G -> Guanina) teniendo el mismo glúcido (Desoxirribosa). Se forma una hebra con cadenas de secuencias de nucleótidos unidas por puentes de hidrógeno formados por bases complementarios de Adenina con Timina (A-T) y de Guanina con Citosina (G-C). [32]

### 2.3.2. GENOMA HUMANO

Un genoma es la colección del ADN completo de un organismo, o sea un compuesto químico que contiene todas las instrucciones genéticas para desarrollar y dirigir las actividades de todo el organismo y que se transmite de generación en generación. El genoma humano tiene 46 cromosomas, agrupados en 23 pares como se puede observar en la Ilustración 10. Estos son heredados de los dos progenitores, de los cuales 22 son autosomas o cromosomas no sexuales y un par es sexual lo que determina el género del individuo [31]. Es decir el cromosoma 23 identifica el sexo en caso de las mujeres los cromosomas en la posición 23 son XX y XY en los hombres.

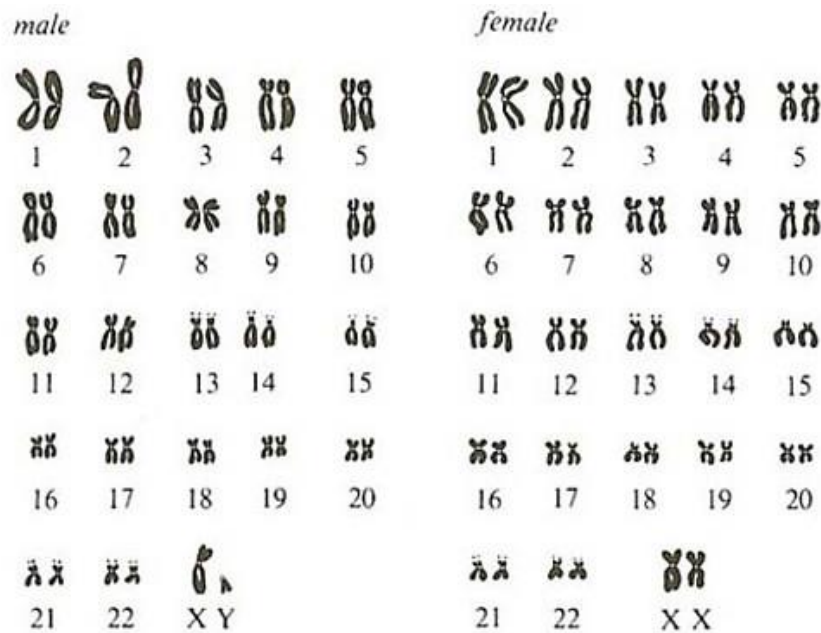


ILUSTRACIÓN 10. GENOMA HUMANO CON LOS 23 PARES DE CROMOSOMAS MASCULINO Y FEMENINO [60]

Muchos organismos no tienen ADN sino tan solo el ARN (Ácido Ribonucleico) como son los virus y plásmidos que realiza la función de transmisión genética a los organismos descendientes. Y esto es en lo que se diferencia el ADN y genoma, el genoma es el ADN y el ARN.

El genoma humano contiene aproximadamente 3.000 millones de pares de bases nucleicos, los cuales se encuentran en los 23 pares de cromosomas dentro del núcleo de todas nuestras células [32]. Cada cromosoma contiene cientos de miles de genes, los cuales tienen las instrucciones para hacer proteínas. Cada uno de los 25.000 genes estimados en el genoma humano produce un promedio de tres proteínas. Estas proteínas realizan las funciones vitales o son parte de ella como los músculos o las enzimas



### 2.3.3. GEN

Un gen es la unidad mínima de información que se guarda en el ADN, mejor dicho en el genoma humano y cuyo objetivo es codificar un producto funcional. Los genes están formados por mezclas de regiones de ADN codificadas conocidas como exones y regiones no codificadas llamadas intrones como se puede apreciar en la Ilustración 11.

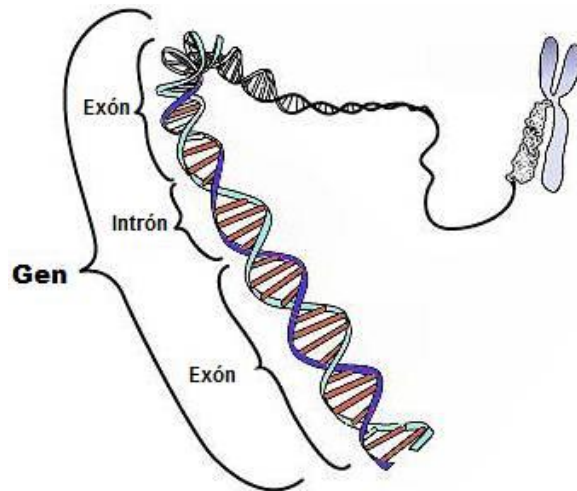


ILUSTRACIÓN 11. GEN COMPUESTO POR EXONES E INTRONES [61]

Un gen, según los expertos, es una serie de nucleótidos que almacena la información que se requiere para sintetizar a una macromolécula que se manifiesta como una característica como sería color de los ojos, del pelo,... etc. Los genes son regiones de ADN que contiene la información necesaria para codificar proteínas, genes reguladores<sup>6</sup> y existen genes que dan lugar a varias proteínas [47]. Existen genes recesivos y dominantes. Para la manifestación de un gen recesivo se debe haber heredado el mismo gen recesivo de los dos progenitores. El conjunto de todos los genes forma el llamado genotipo.

Un gen no es una estructura que se vea sino que se define a nivel funcional. Es una secuencia que va a empezar en algún lugar del ADN y va a terminar en otro. Para conocer un gen se secuencia y se determina la cantidad de los nucleótidos que lo forman y el orden en que se ubican. Muchos de los proyectos nombrados en el apartado 2.2 tienen como objetivo la secuenciación de todos los genes, es decir, del Genoma Humano.

También cabría destacar que existen genes autosómicos dominante y recesivos por los cuales se llegan a heredar enfermedades de los progenitores. En caso de genes autosómicos dominantes si solo uno de los progenitores padece la enfermedad o posee el gen la probabilidad de heredar la enfermedad es de 50%. Pero en caso de genes autosómicos recesivos los dos progenitores deben haber transmitido el gen recesivo.

---

<sup>6</sup> Genes reguladores: son genes que controlan la velocidad de síntesis o creación de los productos en uno o varios genes. [48]



#### 2.3.4. PROTEÍNAS CODIFICADAS A PARTIR DE ADN

Si hay necesidad de utilizar la información codificada en el ADN se realizan copias del fragmento de código donde se encuentra la información y a esta copia se le llama ARN. El proceso de copiar el código ADN se denomina transcripción.

Este ARN también conocido como el ARNm (Ácido Ribonucleico mensajero) se transporta al lugar donde se realiza la síntesis de la proteína conocida como el proceso de Traducción. Este ARN es leído en tripletas o también conocido como codón. Cada codón determina que aminoácido se coloca para formar la proteína. Aunque hay muchos más componentes que intervienen en estos dos procesos se nombran los más relevantes para entender el proceso y se puede ver un resumen en la Ilustración 12.

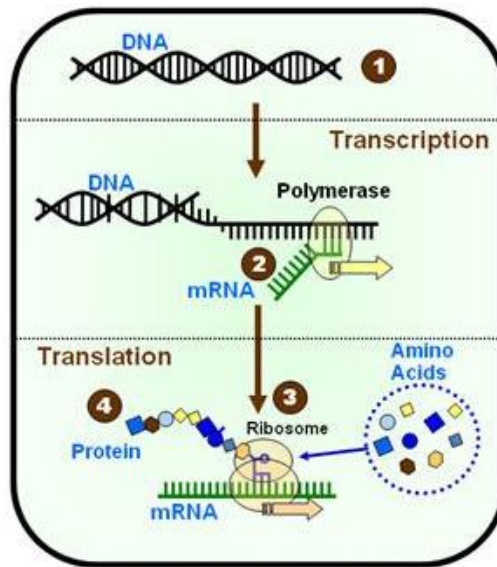


ILUSTRACIÓN 12. DE ADN A PROTEÍNAS: TRANSCRIPCIÓN Y TRADUCCIÓN [62]

Las proteínas, como se ha indicado anteriormente realizan funciones variadas en el organismo humano pero estas están compuestas de aminoácidos. Existen 20 aminoácidos diferentes y cuya combinación es lo que forma proteínas únicas. Aunque este código genético está presente de forma igual en todas las células se utilizan partes específicas de genes dependiendo de qué información es necesaria para realizar las funciones de dicha célula. Por ejemplo las células de estómago realizarán la transcripción de la información relacionada a la digestión.

### 2.3.5. ALELOS

En el genoma humano existen 23 cromosomas como se ha indicado en apartado 2.3.2 y que se heredan de cada progenitor. Es decir, tenemos siempre dos formas iguales o distintas de manifestar un gen o una característica procedentes de cada progenitor. Pero existen los alelos dominantes o recesivos que determinan la manifestación de la característica para una determinada función o rasgo. Los alelos son las 2 posibles formas o versiones que puede tomar un gen. Se puede resumir cual de los dos alelos se manifestará de la siguiente manera:

*Alelo dominante + alelo dominante = fenotipo dominante (alelos homocigóticos)*

*Alelo dominante + alelo recesivo = fenotipo dominante (alelos heterocigóticos)*

*Alelo recesivo + alelo recesivo = fenotipo recesivo (alelos homocigóticos)*

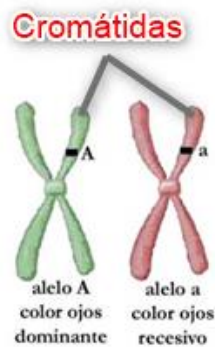


ILUSTRACIÓN 13. ALELOS EN CROMOSOMAS DIPLOIDES [63]

Estos alelos proporcionan una base para poder fijar frecuencias genotípicas. Según el principio de Hardy-Weinberg [48] en una población con alelos sin mutaciones se mantendrá un equilibrio entre las frecuencias de alelos. Ya que sin una mutación solo se llegará a heredar una de las posibilidades anteriores respecto a los alelos. Gracias a esto se puede llegar a calcular las frecuencias de alelos en una población, la frecuencia homocigótica, heterocigótica y la recesiva representada de forma que indica los alelos situados en las dos cromátidas en los cromosomas procedentes de cada progenitor como se puede apreciar en la ilustración 13.

Se debe poder llegar a tener en cuenta las mutaciones ya que su ritmo de aparición es del orden de  $10^{-4}$  a  $10^{-8}$  y el cambio en las frecuencias alélicas será, como mucho, del mismo orden. La validez del principio de Hardy-Weinberg se basa en la asunción de que no se producen nuevas mutaciones. Aunque fuese el caso en general estas mutaciones no dan lugar a expansiones hereditarias grandes debido a sus efectos en los individuos quienes la padecen. [48]

El cálculo de estas frecuencias en varias poblaciones tanto nacionales como continentales es una de las aportaciones del Proyecto Genoma Humano.

### 2.3.6. MUTACIONES

Alguna modificación en la secuencia o forma de ADN puede causar la expresión anormal de uno o más genes, originando un fenotipo patológico. Las enfermedades genéticas causadas por mutación en el genoma pueden ser de distintos tipos pero se pueden clasificar dependiendo del lugar afectado [12]:

- **Mutaciones Cromosómicas:** estas son debido a un cambio en el número o segmentos de cromosomas o una reorganización en el cromosoma. Se consideran todas aquellas que provocan cambios en la estructura del cromosoma [12]. Existe Poliploidía<sup>7</sup>, Aneuploidía<sup>8</sup> o reordenamiento cromosómico (inversión, eliminación, duplicación o translocación en la estructura cromosómica) resumidas en la Ilustración 14.
- **Mutaciones Moleculares o génicas:** por sustitución de bases moleculares (transición o transversión) o los INDELS (por inserción o eliminación de bases). Estos pueden provocar mutaciones silenciosas<sup>9</sup>, mutaciones neutras<sup>10</sup> o mutaciones puntuales<sup>11</sup> [12].

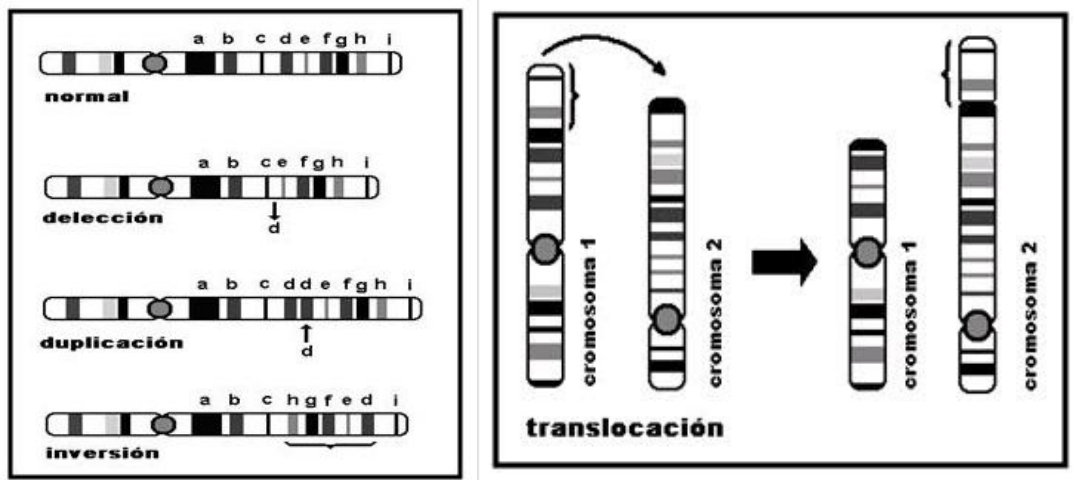


ILUSTRACIÓN 14. MUTACIONES GÉNICAS [64]

<sup>7</sup>Poliploidía: conjuntos adicionales de cromosomas, más de 2 copias del cromosoma. Por ejemplo, poseer 3 cromosomas en lugar de 2 en el par 21.

<sup>8</sup>Aneuploidía: cambio en número de cromosomas. Hay un aumento o disminución en el número total de cromosomas en el genoma y además se realiza una reorganización de los pares. En el genoma humano este tipo de variación resulta la inviabilidad del feto. Pero es muy común en vegetales.

<sup>9</sup> Mutación silenciosa: Cambio en la tercera posición del codón lo que no crea una manifestación fenotípica.

<sup>10</sup> Mutación neutral: el aminoácido insertado es distinto pero no provoca un cambio el comportamiento funcional de la proteína.

<sup>11</sup> Mutación puntual: se cambio un aminoácido y puede llegar a dificultar el futuro funcionamiento y estructura de la proteína.

Las mutaciones pueden ser inducidas o espontáneas. La primera es causada por exposición o por efecto de radiaciones o químicos. Las espontáneas son errores durante la fase o procesos relacionados con el ADN como es la replicación. La diferencia entre una mutación puntual y un SNP es la magnitud de individuos afectados. Es decir, si el polimorfismo de un solo nucleótido se manifiesta en al menos 1 % de la población se considera SNP, sino es considerada una mutación puntual.

### 2.3.7. HAPLOTIPO

Cuando una combinación de alelos son heredados conjuntamente se les conoce como Haplotipo. Esto se diferencia de la herencia normal de los cromosomas ya que en estos alelos se encuentran varios polimorfismos simples de un solo nucleótido, SNP. Existen alrededor de 10 millones de SNP [33] en las poblaciones humanas para los que el alelo del SNP mucho menos común tiene una frecuencia de al menos uno por ciento. Como se ha indicado anteriormente, estos alelos se heredan como un conjunto, un haplotipo como se muestra en la ilustración 15.



ILUSTRACIÓN 15. HAPLOTIPO EN UNA REGIÓN DEL CROMOSOMA [33]

Este concepto se ha utilizado mucho en los estudios genéticos poblacionales. Esto se debe a que si se conoce un SNP asociado a una enfermedad y también el haplotipo asociado a dicho SNP se puede llegar a determinar la posibilidad de padecer dicha enfermedad en un diagnóstico [49]. Los haplotipos proporcionan una mejor comprensión o visión de la herencia genotípica, así como de las enfermedades genéticas. La identificación de haplotipos fue uno de los objetivos principales del Proyecto HapMap.

### 2.3.8. FENOTIPO

Son interpretaciones genotípicas que se manifiestan en forma de características del organismo como el color del pelo o el famoso experimento de los guisantes en la ilustración 16. Estas van asociadas a alelos donde se localiza el gen que lo codifica pero otro punto de vista sería indicar que ese genotipo es una variación genotípica no patogénica. Pero si existen muchas variaciones genéticas patogénicas que se manifiestan como fenotipos y algunos de estos pueden verse afectados por el medio ambiente.

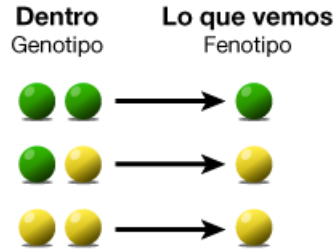


ILUSTRACIÓN 16. GENOTIPO Y SU MANIFESTACIÓN COMO FENOTIPO [65]

### 2.3.9. SECUENCIACIÓN

La secuenciación del genoma consiste en mapear el orden de todas las bases nucleicas presentes en el ADN. Pero esto no proporciona ninguna información por si solo pero puede permitir identificar variantes funcionales asociados a SNPs pertenecientes a enfermedades o variaciones patogénicas.

La secuenciación del ADN nos proporciona las posiciones exactas de los nucleótidos en el genoma humano. Aparte de secuenciar el ADN humano se han secuenciado genomas de muchas plantas, animales y microorganismos. Todo esto permitirá detectar mutaciones, secuenciar fósiles, diagnosis prenatales, pre-implantación e identificar cruces entre especies [34] y proporciona una oportunidad para la biología evolutiva de proponer tratamientos pre-sintomáticos.

En 2000 se secuenció por primera vez casi totalmente el genoma humano después de ello se realizaron múltiples proyecto internacionales para realizar más secuenciaciones<sup>12</sup>. Durante este medio siglo se han estudiados rutas de sistemas biológicos, regulación de expresiones e interconexiones biosintéticas. Pero ahora se ha empezado con la revolución de equipos de investigadores pertenecientes a campos variados como médicos, informáticos, genéticos, bioinformáticos, bioquímicos, físicos y matemáticos para comprender por completo nuestro código genético.

### 2.3.10. MEDICINA DE PRECISIÓN

Según la Sociedad Española de Oncología Médica (SOEM) la medicina de precisión es:

*“Un concepto nuevo que hace referencia a la adaptación del tratamiento médico a las características individuales de cada paciente. Implica que las decisiones referentes al tratamiento o la prevención de enfermedades se tomarán en base a la integración de las características genómicas y moleculares del tumor, la información sobre la situación clínica y los hábitos del paciente.”*

Popularizada en 2015 es la forma de tratar y diagnosticar enfermedades teniendo en cuenta tanto el medio ambiente como el estilo de vida del paciente. Y poder realizar tratamientos seguros y eficientes para poder prevenir enfermedades genéticas. [35]

<sup>12</sup> Se han indicados algunos de ellos en el apartado 2.2.

## CAPÍTULO 3. ENFERMEDADES A TRATAR

### 3.1. DIABETES MELLITUS TIPO 2

La diabetes mellitus, también conocida como diabetes tipo 2, es una enfermedad crónica que solía producirse generalmente en adultos mayores de 50 años, se la conocía como diabetes de los adultos. Sin embargo, en los últimos años su aparición en niños y adolescentes ha sido una causa de preocupación pero muchos la asocian al aumento de la obesidad infantil.

En condiciones normales, la glucosa procede de la digestión de los alimentos y se concentra en el flujo sanguíneo para llegar hasta las células en el cuerpo y transformarse en energía. Esta transformación se realiza por la presencia de una hormona llamada insulina que se produce en el páncreas. Sin embargo, puede llegar el momento en que las células no reaccionen como deberían ante la presencia de insulina. Es lo que se denomina resistencia a la insulina y es el paso anterior a la diabetes tipo 2.

La consecuencia es que la glucosa llega hasta las células pero estas no la absorben del todo y la glucosa se acumula en la sangre, esto provoca que el páncreas produzca mayor cantidad de insulina para lograr que los azúcares sean absorbidos por las células y no se mantengan en el torrente sanguíneo. De este modo se produce un círculo vicioso que es el origen de la diabetes mellitus. La diabetes de tipo 2 provoca niveles elevados de glucosa en la sangre y es uno de los principales factores de riesgo cardiovascular.

Mayoritariamente se asocia a causas características de los pacientes y los factores comunes existentes [14]:

- Obesidad o sobrepeso
- Hiperlipidemia, debido al aumento de los niveles de colesterol en sangre.
- Hipertensión arterial.
- Alimentación inadecuada (dietas hipercalóricas y ricas en grasas saturadas y poli-insaturadas).
- Sedentarismo (se ingieren más calorías de las que se consumen).
- El denominado síndrome metabólico, que se caracteriza por la presencia de tres o más factores de riesgo cardiovascular.

Además, también se indica la existencia de factor genético, ya que los descendientes de personas que tienen este tipo de diabetes tienen una mayor predisposición para desarrollarla. En su fase inicial la diabetes tipo 2 generalmente no produce síntomas y suele ser diagnosticada tras la realización de una analítica clínica rutinaria [14].

Sin embargo, llega un momento en que la glucosa no llega en cantidades suficientes a las células y empieza a acumularse en la sangre y el cuerpo lo manifiesta de las formas resumidas en la ilustración 17.

# DIABETES

## Conoce sus síntomas



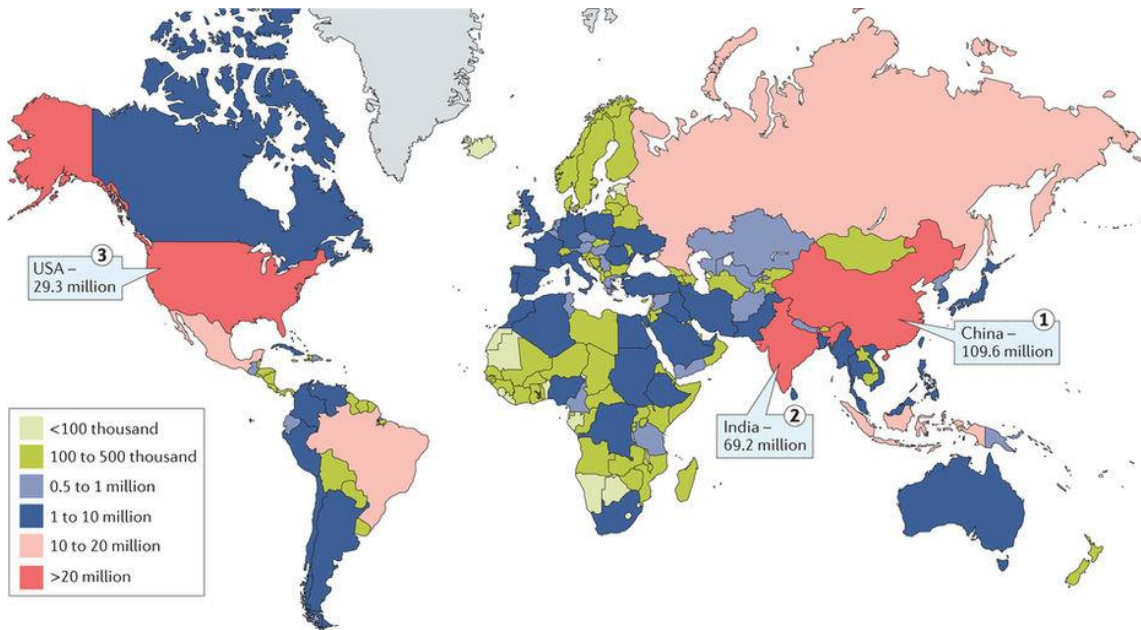
ILUSTRACIÓN 17. SÍNTOMAS DE DIABETES MELLITUS TIPO 2 [66]

Es frecuente que en algunos casos los síntomas se asocien al factor edad y que la existencia de la diabetes mellitus pase desapercibida y continúe su evolución progresiva. Es muy importante, por tanto, que los adultos mayores de 50 años, especialmente si tienen antecedentes familiares, se sometan a una analítica anual de la glucosa en sangre. Ello permitirá la instauración del tratamiento más adecuado en cada caso, con el objetivo de controlar la presencia de azúcares en la sangre. [14]

Ya existen fármacos que están teniendo en cuenta las variantes genéticas para la realización de terapias anticoagulantes orales. Un ejemplo de ello es la Antivitaminas K. En España hay más de 500.000 pacientes (>1% de la población) en tratamiento y prevención de la enfermedad tromboembolia con estos fármacos. [19]

El número de personas que padece diabetes mellitus se ha multiplicado por cuatro en los últimos 30 años, afecta ya a uno de cada once adultos y el 90% de ellos sufre diabetes mellitus de tipo 2 (DMT2). [19] Estas son algunas de las principales conclusiones de una revisión publicada en Nature Reviews Endocrinology que asegura que esta patología es la novena causa de muerte a nivel global.





Nature Reviews | Endocrinology

ILUSTRACIÓN 18. MAPAMUNDI CON LA ESTADÍSTICA DE CASOS DE DIABETES MELLITUS TIPO 2 [15]

### 3.2. TROMBOSIS VENOSA PROFUNDA

La trombosis venosa profunda (TVP) ocurre cuando un coágulo sanguíneo se forma en una de las venas profundas del cuerpo. Esto puede suceder si una vena se daña o si el flujo sanguíneo de una vena se reduce o para. Aunque hay una gran cantidad de factores para desarrollar TVP, dos de los más comunes son experimentar una lesión en la parte inferior del cuerpo y tener una cirugía de cadera o piernas. [36]

Aunque la TVP puede ocurrir en cualquier vena profunda, ocurre más comúnmente en las venas de la pelvis, las pantorrillas o los muslos. Un coágulo sanguíneo puede desplazarse al flujo sanguíneo como se ve en la ilustración 19 y bloquear el flujo de sangre a los pulmones. Aunque un coágulo sanguíneo no se desprenda, puede causar un daño permanente a las válvulas de la vena. Este daño puede conducir a problemas a largo plazo en las piernas, como dolor, inflamación y llagas. [37]

En muchos casos, la TVP ocurre sin síntomas observables y es muy difícil de detectar. Por este motivo, los médicos se enfocan en prevenir el desarrollo de TVP usando diferentes tipos de terapias, dependiendo de las necesidades de un paciente. Si el médico toma las medidas necesarias para prevenir la TVP en caso de tener una fractura grave o una cirugía en las extremidades inferiores.



### Deep Vein Thrombosis (DVT)

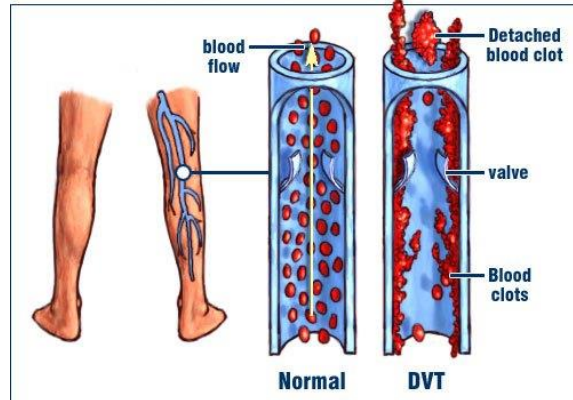


ILUSTRACIÓN 19. TROMBOSIS VENOS PROFUNDA [67]

Varios factores pueden afectar el flujo sanguíneo en las venas profundas y aumentar el riesgo de desarrollar coágulos sanguíneos. Estos incluyen [36]:

- Edad avanzada
- Historial familiar de TVP o embolismo pulmonar
- Tener ciertos tipos de cánceres malignos
- Tener enfermedades en las venas
- Tabaquismo
- Usar píldoras anticonceptivas o terapia hormonal
- Embarazo
- Tener sobrepeso u obesidad
- Heredar un trastorno de coagulación

Las paredes de las venas son lisas. Esto ayuda a que la sangre fluya libremente y se mezcle con agentes presentes naturalmente (anticoagulantes) en la sangre que evitan que las células sanguíneas se coagulen. La sangre que no fluye libremente y no se mezcla con los anticoagulantes tiene más probabilidad de ocasionar coágulos sanguíneos. Por esto es importante observar los signos de TVP en personas que están en reposo, inmobilizadas o que no se pueden mover por períodos prolongados de tiempo. [19]

En ocasiones puede ocurrir el llamado embolismo pulmonar que es un coágulo de sangre que se desprende y viaja por las venas. Esto puede suceder justo después de la formación del coágulo o puede suceder días después. Si el coágulo llega a los pulmones, puede obstruir el flujo de sangre a los pulmones y al corazón. [19]

Un embolismo pulmonar es una emergencia médica grave y puede ocasionar la muerte. En algunos casos puede ser el primer signo de TVP. Los síntomas de embolismo pulmonar incluyen [37]:

- Dificultad para respirar
- Inicio repentino de dolor en el pecho
- Tos
- Escupir o vomitar sangre

La prueba más común de TVP es Ultrasonido dúplex. El ultrasonido usa ondas sonoras de alta frecuencia que hacen eco en el cuerpo, como la tecnología usada para revisar el bienestar de un bebé en gestación. Esto crea una imagen de los vasos sanguíneos. El ultrasonido dúplex combina la tecnología tradicional de ultrasonido con la tecnología Doppler, que genera una imagen de color que muestra la sangre mientras fluye por el cuerpo [19]. El ultrasonido no es invasivo y es indoloro. Se puede repetir regularmente porque no requiere radiación.

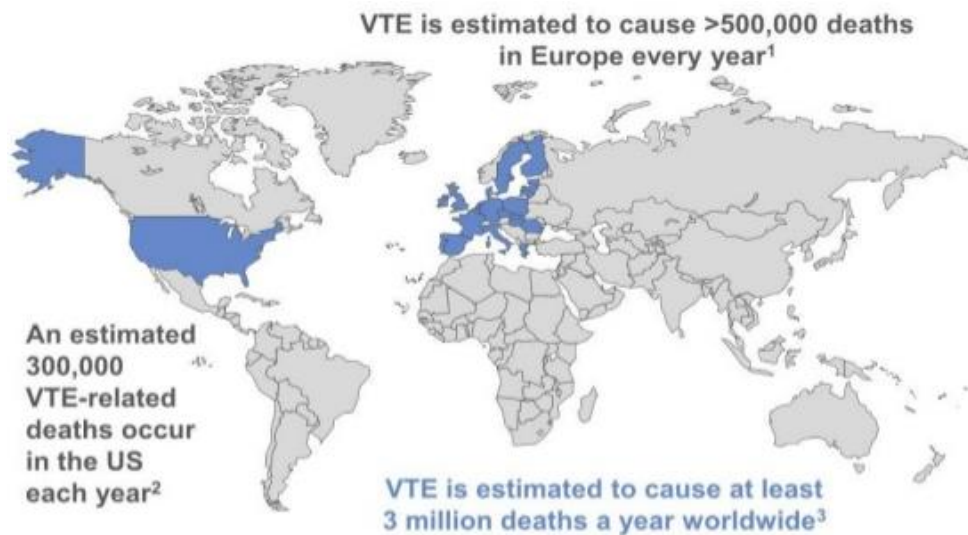


ILUSTRACIÓN 20. TROMBOSIS VENOSA PROFUNDA EN EL MUNDO [68]

## CAPÍTULO 4. REPOSITARIOS TRATADOS

La gran cantidad de datos generados durante las investigaciones y estudios genómicos se guardan en bases de datos para su consulta posterior en busca de información relevante. El objetivo principal de las bases de datos es tener la información almacenada de forma organizada para que sea accesible, actualizada y fácil de manejar.

Según el catálogo NAR (Nucleic Acids Research) existen 1.737 bases de datos biológicas clasificadas en 15 categorías y 41 sub-categorías. En enero de 2018 se publicaron las actualizaciones de las bases de datos entre los que se encuentra el mapeo 3D de las cromatinas. En total se añadieron 88, se descartaron 47 obsoletas y se actualizaron 138 bases de datos con sus nuevas URL's, nuevas descripciones y otros metadatos [3]. Entre las bases actualizadas también se encuentran Ensembl y ClinVar.

La principal razón de selección de estas bases de datos es la presencia de estudios y resultados sobre variaciones de un solo nucleótido, SNPs. Ya que en este trabajo solo se tratan SNPs debido a su gran frecuencia entre las mutaciones y las aportaciones que tiene para la medicina de precisión. Estos SNPs proporcionan un método para poder entender las enfermedades así como sus causas, efectos, propagación y una alternativa potente hacia el desarrollo de medicamentos y tratamientos eficientes en casos tanto generales como puntuales dando lugar tanto a la medicina preventiva como a la medicina de precisión.

Otra de las razones para seleccionar estos repositorios es la relación que tienen entre ellos. El repositorio GWAS Catalog y Ensembl se referencian entre ellos directamente o por medio de artículos referenciados.

Además, todos ellos proporcionan datos estadísticos sobre la variación, su mapeo genómico, fenotipos, datos sobre genes asociados, población de estudio, resultados del estudio entre otros. Pero también proporcionan los resúmenes de los artículos y la referencia junto al acceso a los artículos completos. Casi todos los artículos tratados están en el repositorio *Europe PMC* [44] y de acceso libre. También se consultaron artículos localizados en la sección de artículos perteneciente a la página de *Nature* [45] y otras revistas relacionadas con la genética referenciados desde PubMed. En casos no se tenía acceso al artículo completo pero si se utilizaba la red de la Universidad Politécnica de Valencia se concedía el acceso a dichos artículos.

## 4.1. ENSEMBL

El proyecto Ensembl se inició en 1999, algunos años antes de que se completara el proyecto del genoma humano. Incluso en esa etapa inicial, estaba claro que la anotación manual de 3.000 millones de pares de bases de secuencia no podría ofrecer a los investigadores acceso oportuno a los últimos datos. Por lo tanto, el objetivo de Ensembl era anotar automáticamente el genoma, integrar esta anotación con otros datos biológicos disponibles y poner todo esto a disposición del público a través de la web. [36]

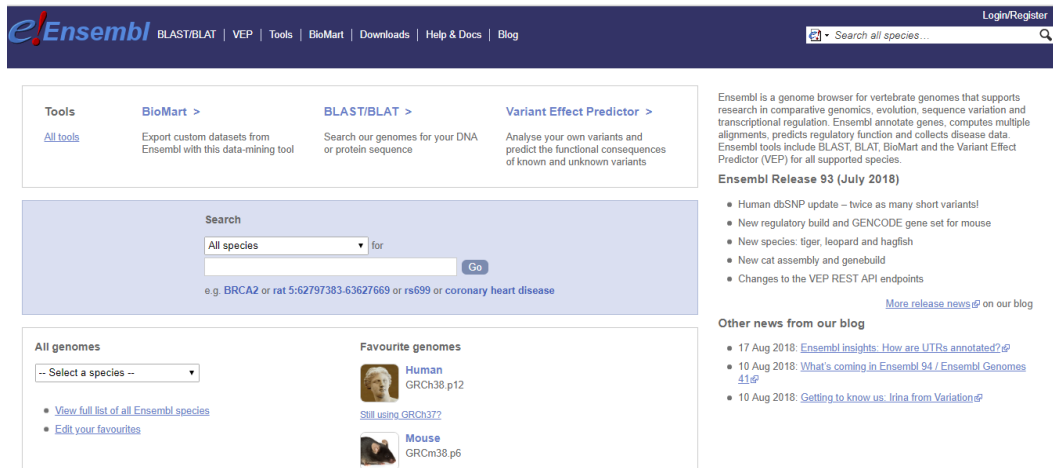


ILUSTRACIÓN 21. REPOSITORIO ENSEMBL [69]

Desde el lanzamiento del sitio web en julio de 2000 y se puede ver en la ilustración 21 la página de inicio de la aplicación web de Ensembl, se han agregado muchos más genomas a Ensembl y el rango de datos disponibles también se ha expandido para incluir genómica comparativa, variación y datos regulatorios [39].

La base de datos de variación de Ensembl almacena áreas del genoma que difieren entre genomas individuales ("variantes") y, cuando están disponibles, información de fenotipos y enfermedades asociadas. Existen diferentes tipos de variantes para varias especies:

- Polimorfismos de un solo nucleótido (SNP)
- Inserciones y / o eliminaciones cortas de nucleótidos.
- Variantes más largas clasificadas como variantes estructurales

Ensembl proporciona la herramienta de datos BioMart por la cual se puede descargar datos genómicos. Los equipos de Comparación, Variación y Regulación son responsables de los datos comparativos, de variación y regulatorios, respectivamente [14]. Además se proporcionan datos separadas en regiones de interés como se muestra en la ilustración 22.

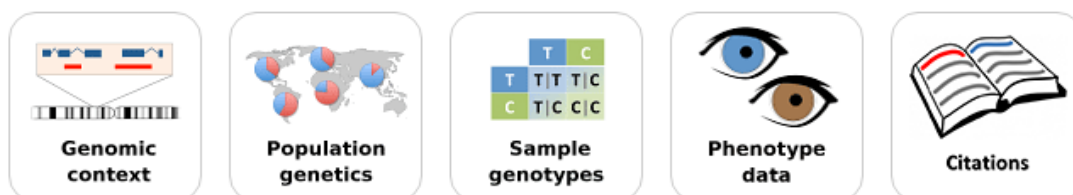
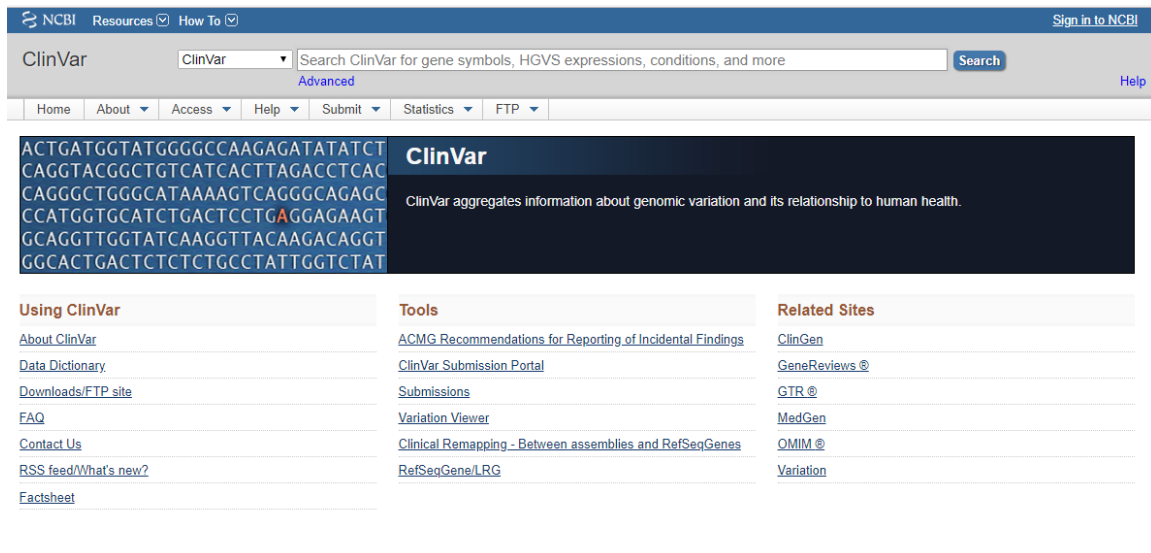


ILUSTRACIÓN 22. REPOSITORIO ENSEMBL REGIONES DE INTERÉS EN LA INTERFAZ WEB [70]

## 4.2. CLINVAR

ClinVar es una base de datos de mutaciones proporciona variantes de cualquier tamaño y región junto con sus significados clínicos, organismos u otros datos de apoyo. Es un archivo público de acceso libre con informes de las relaciones entre variaciones humanas y fenotipos, con evidencia de respaldo y se puede ver la página oficial en la ilustración 23.

ClinVar pertenece al NCBI, que es el Centro Nacional para la Información Biotecnológica y que proporciona acceso a 41 bases de datos (como, por ejemplo, dbVar, dbGaP, PubMed, Gene, OMIM, entre ellas ClinVar). Se trata de una base de datos con gran cantidad de información. Es un socio activo del proyecto ClinGen, proporcionando datos para la evaluación y el archivo de los resultados de la interpretación por reconocidos paneles de expertos y provee guías de práctica. [40]



The screenshot shows the ClinVar website interface. At the top, there is a navigation bar with 'NCBI Resources' and 'How To' menus, and a 'Sign in to NCBI' link. Below this is a search bar with the text 'ClinVar' and a search button. A dropdown menu is open, showing 'Advanced' and 'Help' options. Below the search bar is a navigation menu with links for 'Home', 'About', 'Access', 'Help', 'Submit', 'Statistics', and 'FTP'. The main content area features a large blue box with the text 'ClinVar' and a description: 'ClinVar aggregates information about genomic variation and its relationship to human health.' Below this are three columns of links: 'Using ClinVar' (including About ClinVar, Data Dictionary, Downloads/FTP site, FAQ, Contact Us, RSS feed/What's new?, and Factsheet), 'Tools' (including ACMG Recommendations for Reporting of Incidental Findings, ClinVar Submission Portal, Submissions, Variation Viewer, Clinical Remapping - Between assemblies and RefSeqGenes, and RefSeqGene/LRG), and 'Related Sites' (including ClinGen, GeneReviews®, GTR®, MedGen, OMIM®, and Variation).

ILUSTRACIÓN 23. REPOSITORIO CLINVAR [71]

Las variaciones genómicas y referencias a las secuencias de proteínas se representan siguiendo el estándar de HGVS indicando el cromosoma y la posición de la variación. Por Ejemplo: NC\_000009.12:g.133279294T>G donde NC\_000009.12 es el número único de acceso a la secuencia usada para posicionar la variación, la letra g significa que la secuencia es genómica, 133279294 corresponde a la posición en la secuencia referida y T>G describe el cambio de Timina por Guanina.

ClinVar admite presentaciones de diferentes niveles de complejidad. La presentación puede ser tan simple como una representación de un alelo y su interpretación o tan detallada como proporcionar múltiples tipos de evidencia estructurada o experimental sobre la variación en fenotipos. Se puede ver en la tabla 1 los archivos presentes en 2018 en la base de datos ClinVar.

Categoría	Archivos (3 de Septiembre de 2018)
Archivos totales	701924
Genes representados	30220
Archivos de variaciones únicas	442743
Total de fuentes con aportaciones	1055

TABLA 1. ARCHIVOS PRESENTES EN 2018 EN LA BASE DE DATOS CLINVAR [41]

ClinVar acepta datos obtenidos mediante exámenes clínicos, investigación o conocimiento a base de otros estudios desde distintas partes del mundo. Se puede apreciar en la Ilustración 24 las aportaciones mundiales a la base de datos de ClinVar.

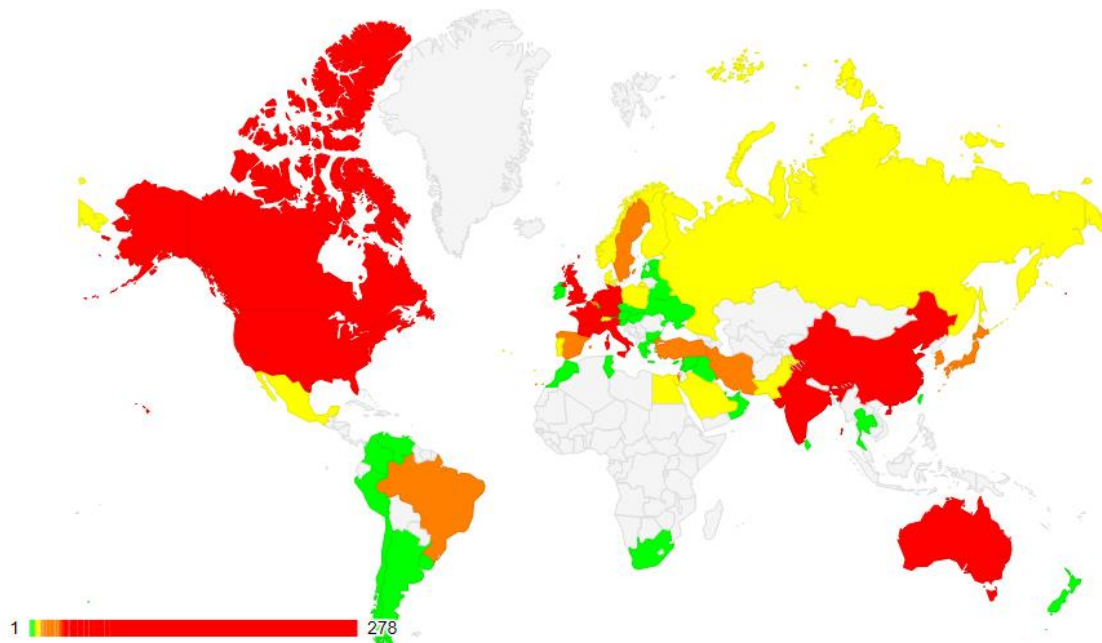
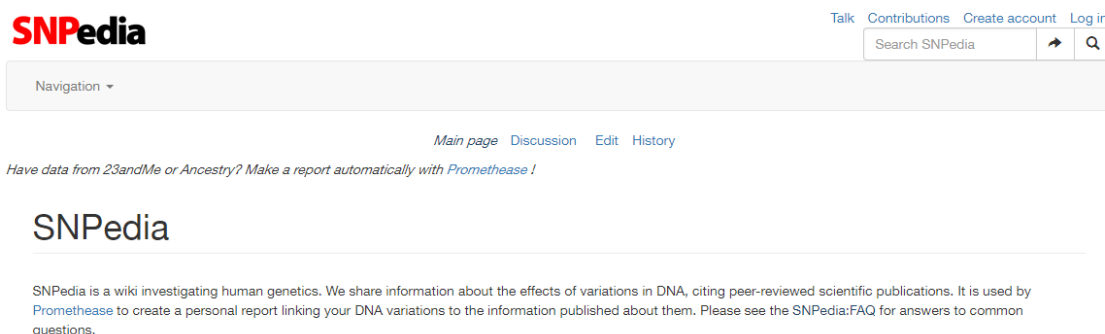


ILUSTRACIÓN 24. LAS APORTACIONES MUNDIALES A LA BASE DE DATOS DE CLINVAR [72]

### 4.3. SNPEDIA

SNPedia es un sitio wiki media semántico que se edita y se actualiza por medios automáticos y manuales. Está diseñado de forma que respalde el análisis y la generación automática de informes por parte del software asociado, al mismo tiempo que mantiene un nivel de legibilidad tanto para usuarios casuales como frecuentes [17]. SNPedia resume los fenotipos, la condición médica y la asociación genealogía de las variantes. Y como el propio nombre indica es una base de datos de polimorfismos simples de un solo nucleótido (SNPs).



**ILUSTRACIÓN 25. REPOSITORIO SNPEDIA [73]**

SNPedia se lanzó para ayudar a desarrollar el potencial del Proyecto del Genoma Humano para conectar la vida diaria y el bienestar. SNPedia se basa en un modelo wiki como se puede ver en la ilustración 25, con el fin de fomentar la comunicación sobre la variación genética y permitir que los miembros de la comunidad interesados lo ayuden a evolucionar para volverse cada vez más relevante [16]. En la tabla 2 se puede ver el estado de los datos [21] disponibles en SNPedia.

Categoría	Archivos (20 de Enero de 2018)
SNPs totales	701924
Genes representados	2042
Temas	136
Total de páginas	260917

**TABLA 2. ARCHIVOS PRESENTES EN 2018 EN LA BASE DE DATOS SNPEDIA [21]**

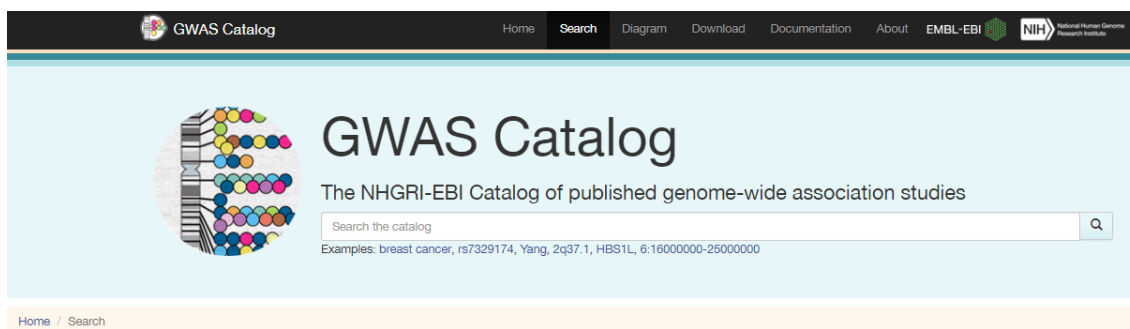
En SNPedia se encuentran SNPs o mutaciones que tienen efectos médicos o genealógicos significativos. El nombre estándar de un SNP registrado oficialmente tiene una 'rs' referencia con un identificador (por ejemplo: rs1051730<sup>13</sup>) y el PMID (Identificador de PubMed) que es un identificador estándar para artículos científicos, según el índice de la Biblioteca Nacional de Medicina de EE. UU.

<sup>13</sup> Esta referencia proviene de la base de datos dbSNP de NCBI que tiene catalogado unos 10 millones de SNPs. [32]

## 4.4. GWAS CATALOG

El Catálogo fue fundado por el NHGRI (Instituto Nacional de Investigación del Genoma Humano) en 2008, debido al rápido aumento en el número de estudios publicados de asociaciones genómicas (GWAS). El catálogo de GWAS proporciona una base de datos consistentes, fácil de buscar y de libre acceso. Con asociaciones de rasgos de SNP publicadas, es accesible para científicos, médicos y otros usuarios.

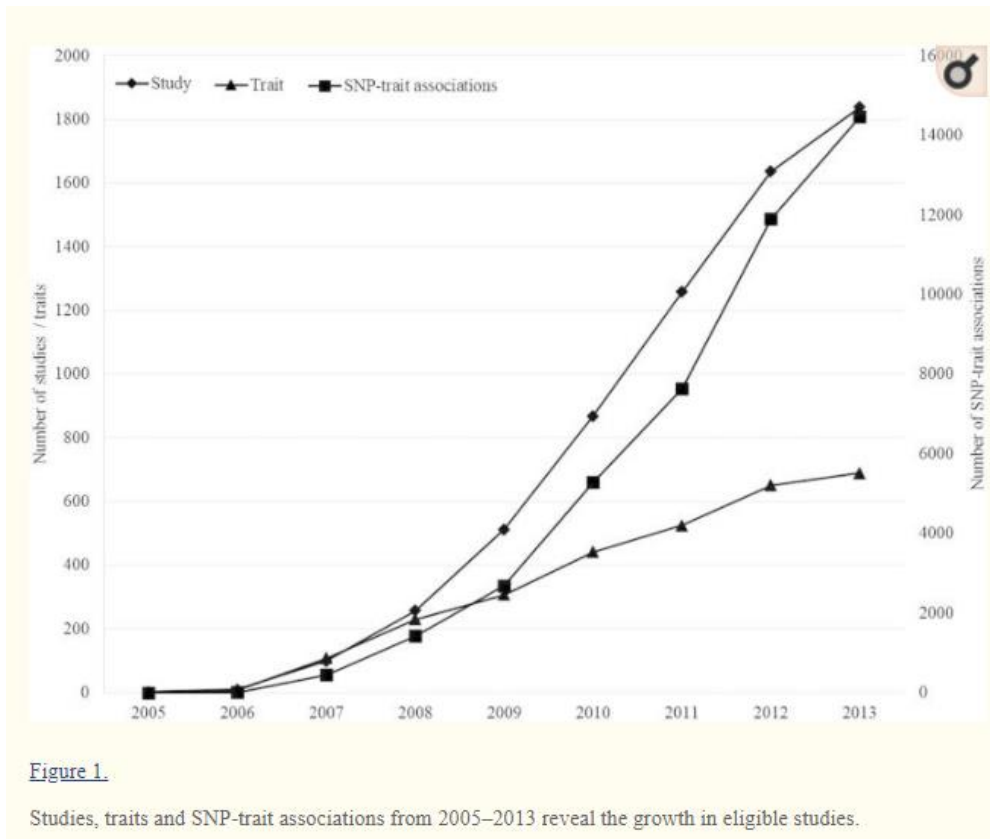
Desde 2010, la entrega y el desarrollo del Catálogo ha sido un proyecto de colaboración entre EMBL-EBI y NHGRI. Un equipo de curadores, todos ellos biólogos moleculares con experiencia, seleccionan los datos y brindan soporte al usuario. El equipo de curación cuenta con el apoyo de desarrolladores de software dedicados y se puede ver el resultado de la página oficial en la ilustración 26. [43]



**ILUSTRACIÓN 26. REPOSITORIO GWAS CATALOG [74]**

En la actualidad, el Catálogo GWAS contiene más de 3,400 publicaciones y más de 62,000 asociaciones exclusivas de Rasgos de SNPs. Más de 10,000 visitantes de todo el mundo acceden al Catálogo GWAS por mes. Las personas que usan el recurso tienen una amplia gama de objetivos: incluyen investigadores, científicos que investigan mecanismos de enfermedad, clínicos que pretenden predecir el riesgo de enfermedad, profesionales de industria farmacéutica que mejoran el proceso de descubrimiento de fármacos y cualquier persona que desee obtener las últimas estadísticas sobre el conocimiento de la enfermedad o datos resumidos de poblaciones particulares de individuos. Se puede apreciar la asociación creciente de los estudios sobre variaciones de un nucleótido simple en la ilustración 27. Así como, GWAS también referencia a otras bases de datos Ensembl, dbSNP y PubMed para enriquecer la información ofrecida.





**XILUSTRACIÓN 27. ESTADÍSTICA GWAS CATALOG 2013 [22]**

## 4.5. ESTRUCTURA DE DATOS UTILIZADA

Las 4 base de datos fuentes de las variaciones con que vamos a trabajar constituyen parámetros distintos para representar las variaciones pero existen parámetros que si ayudan a identificar si se trata de la misma variación y en ese caso fusionar la misma variación en uno. Se juntaron los datos en un único fichero para poder tratarlos obteniendo 29 parámetros para cada variación (Tabla 3).

Los datos procedentes de Ensembl, ClinVar, SNPedia y GWAS Catalog constituyen la siguiente estructura conjunta:

	<b>Descripción</b>	<b>Ejemplo</b>	<b>Procedencia</b>
<b>DBSNP</b>	Identificadores único de entrega proporcionado por dbSNP	rs18726354	dbSNP
<b>CLINVAR_ID</b>	Identificador proporcionado por ClinVar	7413	ClinVar
<b>GRCh37_CHROMOSOME</b>	Nombre del cromosoma donde se inicia la variación.	1	ClinVar o dbSNP
<b>GRCh37_START</b>	Localización de inicio en el genoma de referencia GRCh37	6290728	
<b>GRCh37_END</b>	Localización de fin en el genoma de referencia GRCh37	6290728	
<b>GRCh38_CHROMOSOME</b>	Nombre del cromosoma donde se inicia la variación	1	
<b>GRCh38_START</b>	Localización de inicio en el genoma de referencia GRCh38	22134095	
<b>GRCh38_END</b>	Localización de fin en el genoma de referencia GRCh38	22134096	
<b>RISK_ALLELE</b>	Para los rasgos de la enfermedad, definimos los alelos de riesgo como variantes que corresponden a un OR de enfermedad mayor que uno.	A (Adenina)	
<b>LAST_REVIEWED</b>	Última fecha de revisión	2015	ClinVar
<b>GENOTYPE_1</b>	Es el alelo heredado en una posición del SNP	(C;C)	Ensembl (HapMap o 1000 Genome)
<b>GENOTYPE_1_SUMMARY</b>	Breve descripción de Genotype1		
<b>GENOTYPE_2</b>	Es el alelo heredado en una posición del SNP	(C;T)	
<b>GENOTYPE_2_SUMMARY</b>	Breve descripción de Genotype2		
<b>GENOTYPE_3</b>	Es el alelo heredado en una posición del SNP	(T;T)	
<b>GENOTYPE_3_SUMMARY</b>	Breve descripción de Genotype3		

<b>INITIAL_SAMPLE_SIZE</b>	Tamaño inicial de sujetos en el estudio	<i>12,931 European ancestry cases, 57,196 European ancestry controls</i>	GWAS Catalog
<b>REPLICATION_SAMPLE_SIZE</b>	Número de sujetos en la replicación del estudio	<i>116 Arab ancestry cases, 199 Arab ancestry controls</i>	GWAS Catalog
<b>P-VALUE</b>	Nivel de evidencia de la relación genotipo-fenotipo asignada por el estudio	2,00E-06	GWAS Catalog
<b>OR</b>	Es la medida utilizada para indicar si la presencia o falta de algo está relacionada con la presencia o falta de otro factor [23]	1.1	GWAS Catalog
<b>CI</b>	Es el rango de confianza	[0.016-0.162]	GWAS Catalog
<b>PUBMED_ID</b>	Identificador de la referencia bibliográfica en la base de datos PubMed	29221444	PMID
<b>YEAR</b>	Año de publicación	2017	GWAS Catalog
<b>URL</b>	Enlace al artículo o estudio	<a href="http://www.ncbi.nlm.nih.gov/pubmed/29221444">www.ncbi.nlm.nih.gov/pubmed/29221444</a>	GWAS Catalog
<b>TITLE</b>	El título del estudio o artículo	Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes.	GWAS Catalog
<b>DATABASE</b>	Base(s) de datos de la(s) que se extrae al información	ClinVar	Todas
<b>GEN</b>	Nombre del gen mapeado donde se encuentra la variación	APP	ClinVar
<b>PHENOTYPE</b>	Enfermedades a las que se asocia	<i>TYPE 2 DIABETES</i>	Todas
<b>SIGNIFICADO CLÍNICO</b>	El efecto clínico de la variación	<i>Pathogenic</i>	ClinVar

TABLA 3. PARÁMETROS DE SALIDA DESDE ENSEMBL, CLINVAR, SNPEDIA Y GWAS

# CAPÍTULO 5. METODOLOGÍA SILE

## 5.1. INTRODUCCIÓN

Los modelos conceptuales nos proporcionan una perspectiva comprensiva de los datos. Existen muchas base de datos desde las últimas dos décadas pero que varían en contenido, recursos, infraestructura y calidad. La búsqueda (*Search*) e identificación de información genómica relevante se ha convertido en una tarea que requiere mucho tiempo, experiencia y conocimientos por parte del investigador debido a la variedad o ‘caos’ estructural de los repositorios.

Según los expertos no hay un protocolo o método para buscar e identificar información relevante. Así que, se debe tener en cuenta que si se consultan varias bases de datos la información final tendrá una mejor calidad solo si se realiza el proceso de selección e obtención de forma sistemática.

La metodología SILE (*Search-Identificación-Load-Exploitation*) nos permite realizar todo el proceso de forma sistemática. En este Trabajo fin de máster solo se desarrolla la etapa de *Search e Identification*. Debido a que esta segunda supone un gran esfuerzo y trabajo detallado para tener variaciones genómicas que afectan a un nucleótido con datos y evidencias completas y unificadas.

En el proceso de búsqueda se empieza con la selección de bases de datos fuentes para las variaciones simples de un nucleótido SNPs identificadas para Diabetes Mellitus Tipo 2 y Trombosis Venosa Profunda.

El siguiente paso es la identificación donde se aplican criterios de calidad a los estudios y artículos para clasificar la validez y completitud de los datos. Así como, descartar las variaciones que no son de importancia clínica para este trabajo. Posteriormente, descartar los duplicados, corregir y llenar los valores en las variaciones.

Y por último, completar los 38 campos pertenecientes a los datos asociados a las variaciones para la carga.

## 5.2. SEARCH

Aquí debemos decidir qué medio de información utilizar es decir, el contexto de búsqueda. En nuestro caso se realizar búsqueda de variaciones genómicas relacionadas con la enfermedad de: ***Diabetes Mellitus Tipo 2*** y ***Trombosis Venosa Profunda***.

Y las bases de datos de extracción son Ensembl, SNPedia, ClinVar y GWAS Catalog.

Los datos de descarga fueron proporcionados por la directora experimental Ana León, agrupados por enfermedad y se puede ver un resumen de esos datos en la tabla 4.

	Diabetes Mellitus Tipo 2	Trombosis Venosa Profunda
ClinVar	477	59
Ensembl	23356	896
GWAS Catalog	1281	292
SNPedia	2467	0
Total	27581	1247

TABLA 4. RESULTADOS BÚSQUEDA DIABETES MELLITUS TIPO 2 Y TROMBOSIS VENOSA PROFUNDA

### 5.3. IDENTIFICATION

El objetivo de esta fase es dejar los datos sin redundancias y con la mejor calidad posible. Para ello se han aplicado varios pasos de filtrado a los datos entre los que se encuentra la selección de significado clínico, número de publicaciones asociadas a la variación y unos criterios de calidad. Después de eso se estudiaron las variaciones en busca de inconsistencias Pero para resumir los datos iniciales se pueden clasificar como se muestra en la tabla 5.

	Diabetes Mellitus Tipo 2	Trombosis Venosa Profunda
SNPs	929	205
Estudios / Artículos	2323	426
Genes mapeados	3677	172
Año de Publicación	2007-2018	2009-2017
Filas totales de Variaciones	27581	1247

TABLA 5. RESULTADOS BÚSQUEDA DIABETES MELLITUS TIPO 2 Y TROMBOSIS VENOSA PROFUNDA (\*DESCARTANDO LAS FILAS CON VALORES VACIOS)

#### 5.3.1. CRITERIOS DE CALIDAD

##### 5.3.1.1. SIGNIFICADO CLÍNICO

El significado clínico es de gran relevancia. Los posibles significados clínicos que se encontraron en los datos son:

- ❖ **Benign:** Variación benigna.
- ❖ **Likely benign:** Variación con una alta probabilidad de ser benigna (mayor del 90%).
- ❖ **Uncertain significance:** cuando no se conoce el efecto clínico.
- ❖ **Pathogenic:** Variación patógena, causante de la enfermedad.
- ❖ **Likely pathogenic:** Variación con alta probabilidad de causar la enfermedad (mayor del 90%).
- ❖ **Affects:** Para las variantes que causan un fenotipo sin enfermedad, como la intolerancia a la lactosa.
- ❖ **Conflicting interpretations:** varios artículos o publicaciones asignan a la variación un significado clínica diferente.
- ❖ **Drug response:** Un término general para una variante que afecta la respuesta a un fármaco, no una enfermedad.
- ❖ **Not Provided :** para artículos o estudios sin una interpretación de importancia clínica.
- ❖ **Risk Factor:** cuando existe posible riesgo de que la variación cause la enfermedad
- ❖ **Protector:** cuando la variación previene en cierta medida la aparición de la enfermedad.

De las relevancias clínicas indicadas anteriormente solo se aceptaron variaciones que indicaran precisamente el significado clínico claro de patogénica, con factor de riesgo y protectoras. Pero

las variaciones en las que no se indicaba el significado clínico no se podían descartar. Por ello se rellenó con el valor 'Asociation' estas variaciones.

### 5.3.1.2. NÚMERO DE PUBLICACIONES ASOCIADAS

Otro factor que se tuvo en cuenta fue las publicaciones asociadas a una variación genética. Esto es importante ya si no existen suficientes estudios o artículos que respalde la referencia de una variación esta no es de validez médica. Esto ocurre cuando hay filas asociadas no pertenecientes a artículos o investigaciones.

### 5.3.1.3. CRITERIOS DE CALIDAD

Se aplicaron 5 criterios de calidad a las variaciones relacionados con los datos estadísticos proporcionados en los estudios. Estos son de gran relevancia ya que justifican y son la prueba de las afirmaciones realizadas en los estudios y se descartaron las que no lo cumplían.

- **Sujetos totales > 500:** el número de individuos que fueron sujetos del estudio sumando los casos y controles posteriores debe superar a la cifra de 500. Esto se debe a que estadísticas analíticas evalúan los efectos del tratamiento y los factores de riesgo en resultados específicos y la probabilidad de que los efectos se deban al azar o no [25]. Esta evaluación se basa en la prueba de hipótesis estadísticas y la significancia estadística dependiendo del:
  - **Tamaño** de la muestra, cuando mayor más probable es que se vea la estadística.
  - **Variabilidad** en las características de los sujetos. Cuanto menor es la variabilidad, más fácil es demostrar la significación estadística.
  - **Magnitud** del efecto observado entre grupos. Cuanto mayor es el tamaño del efecto, más fácil es demostrar la significación estadística.
- **OR<sup>14</sup> = 1:** un OR de 1.0 significa que la variante de ADN no afecta las probabilidades de tener la enfermedad. [26]
- **OR - Patogénicas > 1:** los valores superiores a 1.0 indican una asociación estadística entre esa variante y la enfermedad. [26]
- **OR - Protectoras < 1:** Los valores de OR por debajo de 1 indican una menor asociación o riesgo. [26]
- **IC <> 1:** El intervalo de confianza (IC<sup>15</sup>) es la certeza de que un rango (intervalo) de valores contiene el valor verdadero y preciso de una población que se obtendría si el

---

<sup>14</sup> **OR:** es la medida utilizada para calcular la presencia o la ausencia de una propiedad en una población dada. Es la relación entre probabilidades en un grupo en comparación con las probabilidades dentro de otro grupo, para la asociación entre un alelo o genotipo con un fenotipo. Por lo general, los portadores de un alelo o genotipo menos común se comparan con personas con dos copias del alelo más común. Generalmente esta medida se calcula sobre toda la población pero en caso de ser parte de la población la importancia y el efecto depende del valor de intervalo de confianza. [50] [51]

experimento se repitiera. Proporciona información sobre la magnitud y la dirección de un efecto. El 1 indica que no hay diferencia estadística entre comparaciones y no son de nuestro interés.

Con los últimos 3 criterios de calidad de calidad, en el apartado 5.3.1.1, 5.3.1.2, 5.3.1.3 se descartaron 6552 filas relacionadas a variaciones genéticas pertenecientes a Diabetes Mellitus Tipo 2 y 59 de Trombosis Venosa Profunda. En Trombosis Venosa Profunda solo hizo efecto el criterio 5.3.1.1 ya que la falta de los datos no hizo posible efecto de los otros criterios.

Pero cabe destacar que solo se eliminaron unas cuantas filas al aplicar el criterio 5.3.1.3 para las dos enfermedades. Esto se debía a que se utilizan valores estadísticos resultantes de los estudios para valorar la calidad y valor de información que aportan los datos para el diagnóstico clínico significativo pero al faltar muchos de estos datos no se pudieron aplicar estos criterios a las filas asociadas a las variaciones genéticas.

Cabe señalar que hubo casos donde se vieron inconsistencias relacionadas a las variaciones como por ejemplo para un mismo SNP no puede haber dos o más filas con valores estadísticos distintos. Estas y muchas otras inconsistencias en los datos fueron lo que marcaron que la fase de Identificación fuese tan intensa. Este tipo de inconsistencias fueron resueltas uno en uno como se detallando las más significativas en los siguientes apartados.

Esta falta de datos provocó que se aplicaran los filtros de calidad del apartado 5.3.1.1, 5.3.1.2, 5.3.1.3 después de haber completado y revisado las inconsistencias los datos.

---

IC: el intervalo de confianza el resultado estadístico calculado como log en base e de la población estudiada.

### 5.3.2. INCONSISTENCIAS EN LOS DATOS

#### 5.3.2.1. VARIACIONES DE ENFERMEDADES RELACIONADAS

También se observó que había filas de variaciones que apuntaban a un fenotipo distinto del tratado en este trabajo. Cuando se estaban rellenando los valores faltantes en las variaciones se vio que había algunas variaciones de enfermedades relacionadas. Como se puede observar en la Ilustración 28 los fenotipos mapeados en Ensembl perteneciente al SNP 'rs1040196' son distintos a Diabetes Mellitus Tipo 2 pero se encontraban en los datos descargados desde la base de datos Ensembl.

Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	External reference	Reported gene(s)	Associated allele	Statistics
<a href="#">Age-related diseases, mortality and associated endophenotypes</a>	<a href="#">NHGRI-EBI GWAS catalog</a>	16 terms	16 accessions	<a href="#">PMID:27790247</a>	NR	-	p-value: 4.00e-11
<a href="#">Age-related disease endophenotypes</a>	<a href="#">NHGRI-EBI GWAS catalog</a>	9 terms	9 accessions	<a href="#">PMID:27790247</a>	NR	-	p-value: 2.00e-11
<a href="#">Hepatic lipid content in extreme obesity</a>	<a href="#">NHGRI-EBI GWAS catalog</a>	hepatic lipid content measurement, obesity	<a href="#">EFO:0001073</a> , <a href="#">EFO:0006506</a>	<a href="#">PMID:25246029</a>	<a href="#">SUGP1</a>	-	p-value: 3.00e-7
<a href="#">Type 2 diabetes</a>	<a href="#">NHGRI-EBI GWAS catalog</a>	type II diabetes mellitus, Type II diabetes mellitus	<a href="#">EFO:0001360</a> , <a href="#">HP:0005978</a>	<a href="#">PMID:22885922</a>	<a href="#">CILP2</a>	<a href="#">C</a>	p-value: 7.00e-9 odds ratio: 1.13
<a href="#">Type 2 diabetes</a>	<a href="#">NHGRI-EBI GWAS catalog</a>	type II diabetes mellitus, Type II diabetes mellitus	<a href="#">EFO:0001360</a> , <a href="#">HP:0005978</a>	<a href="#">PMID:22885922</a>	<a href="#">CILP2</a>	<a href="#">C</a>	p-value: 3.00e-7 odds ratio: 1.15
<a href="#">LDL cholesterol</a>	<a href="#">NHGRI-EBI GWAS catalog</a>	LDL cholesterol, lipid measurement, low density lipoprotein cholesterol measurement, total cholesterol measurement	<a href="#">EFO:0004195</a> , <a href="#">EFO:0004529</a> , <a href="#">EFO:0004574</a> , <a href="#">EFO:0004611</a>	<a href="#">PMID:19060906</a>	<a href="#">NCAN</a> , <a href="#">CILP2</a> , <a href="#">PBX4</a>	<a href="#">C</a>	p-value: 2.00e-8 beta 0.05 s.c coefficient: decreas

ILUSTRACIÓN 28. VARIACIONES DE ENFERMEDADES RELACIONADAS ENSEMBL [75]

Viendo esta inconsistencia se decidió acceder al artículo perteneciente a esta ocurrencia de variación (PUBMED '25246029'). Y en el resumen del artículo mostrado en la Ilustración 29, se pudo confirmar que se trataba de una variación perteneciente a otro fenotipo pero el estudio había sido realizado sobre sujetos con la enfermedad tratada en nuestro trabajo. Esto obliga a verificar en todos los casos, aunque en los datos descargados indicara el fenotipo estudiado, verificar si las variaciones apuntaban al fenotipo estudiado.



Individuals with type 2 diabetes have an increased risk of developing nonalcoholic fatty liver disease (NAFLD), and NAFLD patients are also at greater risk for developing type 2 diabetes. Although the relationship between type 2 diabetes and NAFLD is highly interconnected, the pathogenic mechanisms linking the two diseases are poorly understood. The goal of this study was to identify genetic determinants of hepatic lipid accumulation through association analysis using histological phenotypes in obese individuals. Using the Illumina HumanOmniExpress BeadChip assay, we genotyped 2300 individuals on whom liver biopsy data were available. We analyzed total bilirubin levels, which are linked to fatty liver in severe obesity, and observed the strongest evidence for association with rs4148325 in *UGT1A* ( $P < 5.0 \times 10^{-93}$ ), replicating previous findings. We assessed hepatic fat level and found strong evidence for association with rs4823173, rs2896019, and rs2281135, all located in *PNPLA3* and rs10401969 in *SUGP1*. Analysis of liver transcript levels of 20 genes residing at the *SUGP1/NCAN* locus identified a 1.6-fold change in expression of the *LPAR2* gene in fatty liver. We also observed suggestive evidence for association between low-grade fat accumulation and rs10859525 and rs1294908, located upstream from *SOCS2* and *RAMP3*, respectively. *SOCS2* was differentially expressed between fatty and normal liver. These results replicate findings for several hepatic phenotypes in the setting of extreme obesity and implicate new loci that may play a role in the pathophysiology of hepatic lipid accumulation.

#### ILUSTRACIÓN 29. VARIACIONES DE ENFERMEDADES RELACIONADAS: RESUMEN ARTÍCULO [27]

En caso de datos pertenecientes a GWAS Catalog sucedió lo mismo y se procedió a realizar el chequeo de todas las variaciones y en este caso se encontró en el rasgo indicada otras enfermedades relacionadas que pueden llegar a provocar o afectan en alguna medida a la enfermedad estudiada pero no llegan a reportar la conexión directa de la variación con las enfermedades de Diabetes Mellitus Tipo 2 o Trombosis Venosa Profunda como se puede ver en la Ilustración 30.

Author	PMID	Study accession	Publication Date	Journal	Title	Reported trait	Association count
Spracklen CN et al.	28334899	GCST004237	2017-02-21	Hum Mol Genet	Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels.	Triglyceride levels	49
Spracklen CN et al.	28334899	GCST004235	2017-02-21	Hum Mol Genet	Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels.	Total cholesterol levels	81
Spracklen CN et al.	28334899	GCST004233	2017-02-21	Hum Mol Genet	Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels.	LDL cholesterol levels	66
He L et al.	27790247	GCST004046	2016-11-03	Front Genet	Pleiotropic Meta-Analyses of Longitudinal Studies Discover Novel Genetic Variants Associated with Age-Related Diseases.	Age-related disease endophenotypes	31
He L et al.	27790247	GCST004045	2016-11-03	Front Genet	Pleiotropic Meta-Analyses of Longitudinal Studies Discover Novel Genetic Variants Associated with Age-Related Diseases.	Age-related diseases, mortality and associated endophenotypes	44

#### ILUSTRACIÓN 30. VARIACIONES DE ENFERMEDADES RELACIONADAS EN GWAS CATALOG [76]

Otra forma de identificar estas variaciones de mapeo erróneo en las descargas es mediante el título de los artículos o estudios. Como se disponía del título de los estudios en los datos se pudo realizar la confirmación directa en los datos pero en casos de ausencia del título o ante la duda se debía consultar el artículo. Para ello se puede consultar el repositorio origen de la variación (Ensembl, GWAS Catalog, SNPedia o ClinVar) o simplemente realizar una consulta directa en la página de PubMed [51] con el identificador asociado al artículo que es un campo siempre presente en los datos descargados y en la ilustración 31 podemos ver un ejemplo de ello.

The screenshot shows a PubMed search interface. The search term is '29221444[luid]'. The search results show a single entry: 'Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes.' The authors listed are Divers J<sup>1</sup>, Palmer ND<sup>2</sup>, Langefeld CD<sup>3</sup>, Brown WM<sup>3</sup>, Lu L<sup>3</sup>, Hicks PJ<sup>4</sup>, Smith SC<sup>4</sup>, Xu J<sup>4</sup>, Terry JG<sup>5</sup>, Register TC<sup>6</sup>, Wagenknecht LE<sup>7</sup>, Parks JS<sup>8</sup>, Ma L<sup>9</sup>, Chan GC<sup>9</sup>, Buxbaum SG<sup>10</sup>, Correa A<sup>11</sup>, Musani S<sup>11</sup>, Wilson JG<sup>12</sup>, Taylor HA<sup>13</sup>, Bowden DW<sup>4</sup>, Carr JJ<sup>5</sup>, Freedman BI<sup>9</sup>.

ILUSTRACIÓN 31. ARTÍCULO DE ENFERMEDADE RELACIONADA [77]

En este caso se descartaron 2581 filas de variaciones genéticas pertenecientes a Diabetes Mellitus Tipo 2 debido a que estaban mapeando otro fenotipo o fenotipos relacionados con la enfermedad tratada. En caso de Trombosis Venosa Profunda todos los artículos asociados trataban de este fenotipo.

### 5.3.2.2. AFECTADO POR GENES DE PROXIMIDAD

Se encontró el caso de que había varios genes reportados para un solo SNP (rs11043007). Para que una variación afecte a más de un gen ésta debe ser muy grande (no es el caso porque estamos trabajando con SNPs) o que los genes estén solapados (esto tampoco es posible). Para salir de esta incertidumbre se verificó la información la base de datos fuente y constaba con los dos genes reportados, como se puede ver en la Ilustración 32.

The screenshot shows the Ensembl variant page for rs11043007. The variant is located on Chromosome 11:2183058 (forward strand). The most severe consequence is 'intergenic variant'. The variant is associated with 'Type 2 diabetes' (p-value: 1.00e-7, odds ratio: 1.19). The reported gene is ASCL2\_TH.

Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	External reference	Reported gene(s)	Associated allele	Statistics
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001360, HP:0005978	PMID:25102180	ASCL2_TH	G	p-value: 1.00e-7 odds ratio: 1.19

ILUSTRACIÓN 32.SNP LOCALIZADO FUERA DEL GEN EN ENSEMBL [78]

Para localizar cual de los dos estaba mapeado en esa posición se consultó la base de datos dbSNP, Ilustración 33 [52] y en casos resultó ser que ninguno de los dos genes o uno de ellos estaba mapeados con ese SNP (Ilustración 34), uno de esos genes estaba localizado en el flanco izquierdo y otro en el derecho.

ILUSTRACIÓN 33.SNP LOCALIZADO EN DBSNP [79]

ILUSTRACIÓN 34. SNP LOCALIZADO FUERA DEL GEN [80]

Por ello la primera suposición que se hizo fue que estos estaban afectados por proximidad y en este trabajo solo se trata con variaciones de nucleótidos simples. Se encontró con variaciones intergénicas con mapeo en otros genes también y se procedió a realizar su eliminación. Se descartaron 1001 filas de variaciones genéticas pertenecientes a Diabetes Mellitus Tipo 2.

### 5.3.2.3. ESTUDIO DE EFECTO SNP POR SNP

En los datos de Trombosis Venosa Profunda se encontró entre los datos con un gen 'NRxNR' pero al consultar la fuente de repositorio se encontró que se trataba de un estudio sobre la interacción gen con gen en la enfermedad de Trombosis Venos Profunda. Como se puede observar en la Ilustración 35.

Author	PMID	Study accession	Publication Date	Journal	Title	Reported trait	Association count
Greliche N et al.	23509962	GCST001913	2013-03-20	BMC Med Genet	A genome-wide search for common SNP x SNP interactions on the risk of venous thrombosis.	Venous thromboembolism (SNP x SNP interaction)	37

SNP	RAF	p-value	OR	Beta	CI	Region	Location	Functional class	Reported gene(s)	Mapped gene(s)	Reported trait	Publication	Study accession
rs1364505-A x rs1204660-A	2	1.8 x10 <sup>-9</sup>			[NR]	7q32.3 x 20q11.22	7:132345252 x 20:35405288	intron_variant x intron_variant	NR x NR	PLXNA4 x UQCC1	Venous thromboembolism (SNP x SNP interaction)	Greliche N (PMID: 23509962), 2013	GCST001913

ILUSTRACIÓN 35. ESTUDIO VARIACIÓN CON INTERACCIÓN GEN POR GEN. [81]

Estos estudios empezaron (Ilustración 36) después de que estudiaran la interacción del ambiente con la variación genómica en algunas enfermedades conocido como estudios 'SNP-by-environment'. Pero en este trabajo solo se tiene en cuenta las variaciones genómicas que afectan a un solo SNP y por ello se descartaron estos tipos de estudios. Se descartaron 130 filas de variaciones genéticas pertenecientes a Trombosis Venosa Profunda.

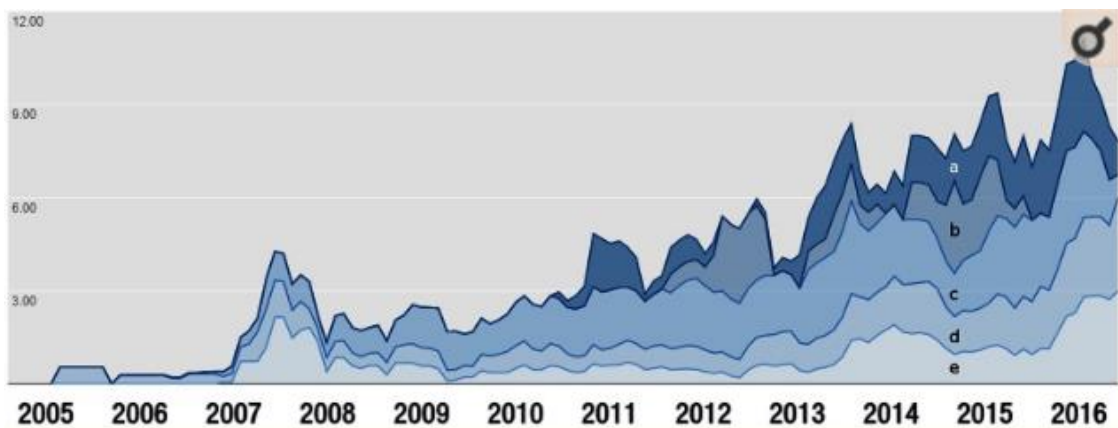


Figure 2.

Increasing complexity of GWAS studies over time (A) number of SNP-by-environment interaction studies, (B) number of SNP-by-SNP interaction publications, (C) number of traits per publication, (D) number of ancestry categories each GWAS publication analyzed and (E) number of GWAS analyses per publication. Values were normalized to provide equal weighting to each category.

ILUSTRACIÓN 36. EVOLUCIÓN DE ESTUDIOS Y ARTÍCULO EN GWAS [28]

#### 5.3.2.4. FALTA DE SUJETOS DE ESTUDIO

En casos no se indicaba el número de sujetos del estudio en los datos finales pero al revisar el artículo asociado se pudo encontrar esta información como se puede apreciar en la Ilustración 37. Estos valores se rellenaron a mano en los datos finales para poder aplicar los criterios de calidad indicados en el apartado 5.3.1 en su totalidad.

NCBI Resources How To

PubMed 25755232[uid]

Format: Abstract

Eur J Endocrinol. 2015 May;172(5):595-601. doi: 10.1530/EJE-14-0805. Epub 2015 Mar 9.

**Habitual coffee intake, genetic polymorphisms, and type 2 diabetes.**

Lee JK<sup>1</sup>, Kim K<sup>1</sup>, Ahn Y<sup>1</sup>, Yang M<sup>1</sup>, Lee JE<sup>2</sup>.

Author information

**Abstract**

**BACKGROUND:** The association between coffee intake and type 2 diabetes may be modulated by common genetic variation.

**OBJECTIVE:** The purpose of this study was to examine the association between habitual coffee intake and the risk of type 2 diabetes and to determine whether this association varied by genetic polymorphisms related to type 2 diabetes in Korean adults.

**DESIGN AND METHODS:** A population-based cohort study over a follow-up of 4 years was conducted. A total of 4077 Korean men and women aged 40-69 years with a normal glucose level at baseline were included. Coffee intake was assessed using a validated food frequency questionnaire, and incident type 2 diabetes or prediabetes was defined by oral glucose tolerance test or fasting blood glucose test. The genomic DNA samples were genotyped with the Affymetrix Genome-Wide Human SNP Array 5.0, and nine single-nucleotide polymorphisms related to type 2 diabetes in East Asian populations were extracted.

ILUSTRACIÓN 37. FALTA SUJETOS DE ESTUDIO [82]

#### 5.3.2.5. FALTA DE VALORES ESTADÍSTICOS PERTENECIENTES A SNPEDIA

En los datos procedentes de SNPedia se vio la ausencia de datos estadísticos. Pero al acceder a SNPedia se encontraban los datos de la forma indicada en la Ilustración 38. En este caso se puede ver que SNPedia está referenciándose a datos correspondientes a otro repositorio, GWAS. Y en casos se encontró valores estadísticos faltantes en las descripciones o resúmenes que ofrece SNPedia para cada artículo asociado. Estos valores se rellenaron a mano en los datos finales para poder aplicar los criterios de calidad indicados en el apartado 5.3.1 en su totalidad.

GWAS snp	
PMID	[PMID 23945395]
Trait	Type 2 diabetes
Title	Genome-wide association study identifies three novel loci for type 2 diabetes.
Risk	C
Allele	
P-val	5E-7
Odds	1.12 [1.07-1.17]
Ratio	

ILUSTRACIÓN 38. FALTA DE VALORES SNPEDIA [83]

### 5.3.2.6. CITACIONES

En casos se encontró con la falta de todos los valores estadísticos y al consultar el artículo se pudo apreciar que se trataba de una referencia o asociación a otro artículo o estudio para respaldar o argumentar alguna hipótesis. En este caso se descartaron 7075 filas de variaciones genéticas pertenecientes a Diabetes Mellitus Tipo 2 y 409 de Trombosis Venosa Profunda.

### 5.3.2.7. POBLACIÓN Y FASES DE ESTUDIOS

Se encontró con el caso de que había varios valores estadísticos pertenecientes a la misma variación y el mismo artículo.

Al consultar la fuente de datos se vio que se trataba de un estudio en varias poblaciones. En este caso se procede a encontrar el valor correcto consultando el artículo perteneciente y por lo general se escoge la que referencia a estadística estándar, es decir la global.

rs1470579- C	0.3042	1 x10 <sup>-45</sup>	0.13 unit increase	[0.11-0.15]	3q27.2	3:185811292	intron_variant	IGF2BP2	IGF2BP2	Type 2 diabetes	Zhao W (PMID: 28869590), 2017
rs1470579- C	NR	2 x10 <sup>-24</sup> (Japanese)	1.176043	[1.13996358956295- 1.21326535811954]	3q27.2	3:185811292	intron_variant	IGFBP2	IGF2BP2	Type 2 diabetes	Imamura M (PMID: 26818947), 2016
rs1470579- C	0.5	2 x10 <sup>-19</sup> (South Asian, East Asian, Europeans)	1.08	[1.05-1.09]	3q27.2	3:185811292	intron_variant	IGF2BP2	IGF2BP2	Type 2 diabetes	Saxena R (PMID: 23300278), 2013
rs1470579- C	0.4571	4 x10 <sup>-14</sup> (South Asian)	0.1264 unit increase	[0.094-0.159]	3q27.2	3:185811292	intron_variant	IGF2BP2	IGF2BP2	Type 2 diabetes	Zhao W (PMID: 28869590), 2017
rs1470579- C	0.34	5 x10 <sup>-14</sup>	1.19	[1.14-1.24]	3q27.2	3:185811292	intron_variant	IGF2BP2	IGF2BP2	Type 2 diabetes	Hara K (PMID: 23945395), 2013

ILUSTRACIÓN 39. POBLACIÓN Y SUBPOBLACIÓN [84]

Se encontraron distintos *p-values* para el mismo artículo (ilustración 39) y SNP, se puede ver en la Ilustración 40 que hay varios valores para el mismo SNP y artículo. Se revisó el artículo correspondiente encontrándose con las distintas fases que existieron en el estudio realizado y posteriormente se descartaron todos menos los valores de la fase final.

Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:2888590#	SLC30A8	I	p-value: 2.00e-7 beta coefficient: 0.0955 unit decrease
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:17483248#	SLC30A8	C	p-value: 5.00e-8 odds ratio: 1.12
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:23845395#	SLC30A8	C	p-value: 5.00e-7 odds ratio: 1.12
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:27189021#	SLC30A8	C	p-value: 1.00e-11 odds ratio: 1.12
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:10056611#	SLC30A8	-	p-value: 7.00e-6 odds ratio: 1.12
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:17483248#	SLC30A8	C	p-value: 5.00e-8 odds ratio: 1.12
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:17483248#	SLC30A8	C	p-value: 5.00e-8 odds ratio: 1.12
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:2888590#	SLC30A8	I	p-value: 2.00e-22 beta coefficient: 0.1113 unit decrease
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:26818947#	SLC30A8	C	p-value: 9.00e-13 odds ratio: 1.12039
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:19401414#	SLC30A8	C	p-value: 2.00e-14 odds ratio: 1.22
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:17283876#	SLC30A8	C	p-value: 6.00e-6 odds ratio: 1.18
Type 2 diabetes	NHGRI-EBI GWAS catalog	type II diabetes mellitus, Type II diabetes mellitus	EFO:0001380# HP:0005978#	-	PMID:2888590#	SLC30A8	I	p-value: 4.00e-42 beta coefficient: 0.109 unit decrease

ILUSTRACIÓN 40. ENSEMBL DISTINTOS P-VALUES [85]

En este caso se descartaron 1937 filas de variaciones genéticas pertenecientes a Diabetes Mellitus Tipo 2 y 152 de Trombosis Venosa Profunda.

### 5.3.3. CRITERIOS DE CALIDAD CON LOS DATOS COMPLETOS

Como se ha visto en el apartado anterior se han rellenado datos faltantes y por ello se procedió a aplicar otra vez los criterios de calidad indicados en el apartado 5.3.1.

En este caso se descartaron 7332 filas de variaciones genéticas pertenecientes a Diabetes Mellitus Tipo 2 y 342 de Trombosis Venosa Profunda.

## 5.4. PREPARACIÓN ANTERIOR A LA CARGA

Después de tratar las inconsistencias encontradas en los datos descargados se puede proceder a realizar la carga. Pero la estructura final esperada para cada variación no es la adecuada para la base de datos de carga. Estos datos son fruto de las descargas de variaciones de los 4 repositorios pero el último paso en la fase de identificación será preparar los datos acorde con el modelo conceptual perteneciente al repositorio de carga (ANEXO 1).

Al final se realizaron los siguientes cambios en las variaciones finales acorde al modelo conceptual para su posible carga en la base de datos con los 38 campos indicados en el ANEXO 1:

- Se añadieron unos 19 parámetros nuevos
- Se descartaron 10 parámetros
- Se modificaron 3 (en nombre o en formato)

Para los últimos cambios indicados se consultaron las bases de datos tanto realizando descargas como consultando la web.

En la consulta manual por web se buscaron los campos como el alelo de referencia, alelo alternativo, frecuencia de referencia, frecuencia del gen alternativo, frecuencia del genotipo 1,



frecuencia del genotipo 2, frecuencia del genotipo 3, nombre de la variante, referencias del artículo y población. Se puede consultar el Anexo 2 para ver las correspondientes capturas.

Otros datos se pudieron descargar desde las bases de datos como la fecha de creación, fecha de la última modificación, nombre del genoma, sinónimos del genoma, identificador del gen, identificador NG, identificador NM e identificador NP. Estos fueron proporcionados por la directora experimental Ana León.

Y por último se juntaron las filas con los mismos datos pero pertenecientes a diferentes repositorios, es decir un *merge* de los datos indicando en el campo de bases de datos todos los repositorios que tenían estos datos. Lo que provocó la eliminación de 10 filas en caso de Diabetes Mellitus Tipo 2.

Con esto se concluye la fase de identificación quedando con resultados indicados en la tabla 6.

	<b>Diabetes Mellitus Tipo 2</b>	<b>Trombosis Venosa Profunda</b>
<b>SNPs</b>	1093	155
<b>Estudios / Artículos</b>	128	19
<b>Genes mapeados</b>	361	116
<b>SNPs únicos</b>	471	135
<b>Filas totales de Variaciones</b>	27581	1247

**TABLA 6. RESULTADOS FINALES**



## CAPÍTULO 6. CONCLUSIONES Y TRABAJOS FUTUROS

La Tesis de Máster que aquí concluyo me ha proporcionado una serie de conclusiones muy interesantes como colofón de mis estudios de Máster. Procedo a desarrollarlos brevemente a continuación.

En primer lugar, la aplicación de un proceso sistemático para la gestión de datos genómicos orientado a la identificación de variaciones genómicas para las enfermedades tratadas me ha permitido proyectar mis conocimientos en Ingeniería de Sistemas de Información sobre un dominio apasionante que es el del entendimiento del genoma humano con aplicaciones clínicas en la Medicina de Precisión y de esta forma tener una primera conexión con este moderno dominio.

El conocimiento y uso de un método concreto para esa gestión completa de datos (el método SILE) y en particular su fase de identificación precisa de variaciones para conectar la búsqueda con la carga, me ha permitido enfrentarme a la complejidad que tiene la gestión eficiente de la información genómica. Se ha experimentado la demanda directa de gestión correcta del “Big Data” genómico.

Se ha visto que la automatización de procesos para el manejo de estos datos genómicos requiere mucho tiempo y esfuerzo para generar entornos de trabajo factibles y compatibles. He aprendido que debemos empezar a organizar y estandarizar los repositorios genómicos entendiendo su estructura e importancia para evitar que llegue el momento en que el parabólico aumento en los datos genómicos no sea para nada manejable debido a la variedad e incompatibilidad de los repositorios.

También he aprendido a determinar qué aspectos garantizan la calidad de los datos seleccionados, y me he enfrentado a dos casos reales donde con un tremendo esfuerzo he podido aprender a seleccionar de un conjunto inicial de miles de variaciones candidatas, cuáles son las realmente relevantes a efectos de ese diagnóstico clínico fiable que constituye el objetivo final fundamental del trabajo realizado. La visión es poder llevar a cabo diagnósticos directos y proporcionar información relevante sobre las posibles enfermedades genéticas y proponer prevenciones y si hace falta tratamiento personalizado para poder tratar la enfermedad.

En un ambiente de Ciencia de Datos, y de gestión de “Big Data”, he podido comprobar la dificultad de acceder a informaciones que, a pesar de tener el mismo objetivo, aparecen en distintos formatos en diversas fuentes de datos, presentando problemas de falta de consistencia, heterogeneidad, fiabilidad variables, en evolución constante... Todo ello especialmente problemático en un entorno como el del diagnósticos clínico desde una perspectiva genómica, que exige una precisión y una fiabilidad máxima.

Por último, he comprobado que mi trabajo solo visualiza la punta del iceberg que la completa y correcta gestión de la información genómica requiere y requerirá en los próximos años. Solo la

identificación conforma un problema enorme, pero por delante tiene la selección de fuentes de datos relevantes entre centenares de repositorios y bases de datos candidatas, y por detrás el diseño e implementación de los mecanismos de carga e interpretación de resultados que haga viable en ambientes clínicos la explotación en un sistema sanitario de última generación de toda esta información.

Con el horizonte de llegar a entender y manipular el genoma humano en última instancia, esta Tesis de Máster pone su granito de arena en la construcción de un cuerpo de conocimiento asociado a nuevos dominios de I+D englobados bajo la etiqueta del “Diseño de Sistemas de Información Genómicos”, o de la “Ciencia de Datos Genómicos”, líneas de I+D claves en el Centro de I+D en Métodos de Producción de Software -PROS- donde he desarrollado mi trabajo, y donde estaría encantada de poder seguir desarrollando todo lo aprendido, centrado en asegurar la fiabilidad de los datos en función de los continuos descubrimientos que en el ámbito genómico aparecen día a día, y en aplicar al diagnóstico genómico de última generación los resultados que presento en mi trabajo.

Durante el desarrollo de este proceso han surgido ciertas ideas que podrían tener aportación en el dominio de Ciencias de Datos Genómicos. Y para terminar se plantean los siguientes trabajos relacionados con la fase de identificación de los datos:

- Programa de aplicación automática de criterios de calidad.
- Programa de interpretación de lenguaje natural para identificar variaciones de fenotipos relacionados.
- Programa de autocompleto de los datos faltantes siguiendo las trazas utilizadas en este trabajo para la búsqueda de los datos faltantes.
- Una posible combinación de los 3 programas anteriores para tener un programa de filtrado automático y de esta forma evitar posibles errores humanos cometidos durante el filtrado de datos.

Si se quiere incorporar más repositorios habría que realizar estudios parecidos para otros repositorios a integrar y manejar los casos de inconsistencias que vayan apareciendo.

## BIBLIOGRAFÍA

1. Información relacionada a diabetes mellitus tipo 2 :  
<http://www.revistanefrologia.com/es-publicacion-suplementosextra-articulo-genetica-diabetes-mellitus-X2013757511002452>
2. Genoma Humano : [https://es.wikipedia.org/wiki/Proyecto\\_Genoma\\_Humano](https://es.wikipedia.org/wiki/Proyecto_Genoma_Humano)
3. NAR , *Nucleic Acids Research*:  
<https://academic.oup.com/nar/article/46/D1/D1/4781210>
4. Tesis Doctoral de J. F. Reyes : <https://riunet.upv.es/handle/10251/99565>
5. Proyecto Genoma Humano :  
[https://es.wikipedia.org/wiki/Proyecto\\_Genoma\\_Humano](https://es.wikipedia.org/wiki/Proyecto_Genoma_Humano)
6. ENCODE, *Encyclopedia of DNA Elements* : <https://es.wikipedia.org/wiki/ENCODE>
7. [https://www.ub.edu/web/ub/es/menu\\_eines/noticies/2015/09/047.html?](https://www.ub.edu/web/ub/es/menu_eines/noticies/2015/09/047.html?)
8. Fuente de Imagen de ADN : <https://www.cancer.gov/espanol/cancer/causas-prevencion/genetica>
9. Fuente Imagen Genoma Humano : <http://sciencedoing.blogspot.com/2013/01/human-male-degenerative-sex.html>
10. Fuente Imagen ADN a proteínas :  
<http://www.monsantoglobal.com/global/py/productos/pages/transcripcion.aspx>
11. Información relacionada a Genes : <http://biologiabi.blogspot.com/2011/06/14-cromosomas-genes-alelos-y-mutaciones.html>
12. Información relacionada a Mutaciones :  
<https://es.wikipedia.org/wiki/Mutaci%C3%B3n>
13. Página oficial de SOEM : <https://www.seom.org/es/informacion-sobre-el-cancer/ique-es-la-medicina-de-precision>
14. Información relacionada a diabetes mellitus tipo 2 :  
<https://www.sanitas.es/sanitas/seguros/es/particulares/biblioteca-de-salud/diabetes/son004004wr.html>
15. Fuente imagen estadística mundial Diabetes Mellitus Tipo 2 :  
<http://laesalud.com/hackathonsalud/2018/hackathon-salud/adultos-afectados-diabetes-mellitus-mundo/>
16. Información relacionada a trombosis venosa profunda :  
<https://orthoinfo.aaos.org/es/diseases--conditions/trombosis-venosa-profunda-deep-vein-thrombosis/>
17. Fuente Imagen trombosis venosa profunda :  
<https://www.slideshare.net/GamalAgmy/updates-in-venous-thromboembolism>
18. Información relacionada a trombosis venosa profunda :  
<https://www.businesswire.com/news/home/20160323005792/en/Boehringer-Ingelheim-launches-RE-COVERY-DVTPE%E2%84%A2-global-observational>
19. Información relacionada a trombosis venosa profunda :  
[https://es.wikipedia.org/wiki/Diabetes\\_mellitus\\_tipo\\_2](https://es.wikipedia.org/wiki/Diabetes_mellitus_tipo_2)
20. Artículo *Genotipos frecuentemente asociados a trombofilia* :  
<http://www.scielo.org.co/pdf/bio/v34n1/v34n1a16.pdf>

21. Datos estadísticos pertenecientes al repositorio SNPedia :  
<https://www.snpedia.com/index.php/SNPedia:FAQ>
22. Fuente imagen de estadística GWAS : <http://blog.goldenhelix.com/goldenadmin/the-10th-anniversary-of-gwas/>
23. Información relacionada a Odds Ratio: [https://en.wikipedia.org/wiki/Odds\\_ratio](https://en.wikipedia.org/wiki/Odds_ratio)
24. Información relacionada a significados clínicos :  
<https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/>
25. Información relacionada a criterios de calidad :  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3506143/>
26. Información relacionada con los campos utilizados en la base de datos :  
<https://www.snpedia.com/index.php/Glossary>
27. Artículo de enfermedades relacionadas:  
<http://europepmc.org/articles/PMC4370808;jsessionid=A3AE577F7963EC2CBF9EC105C2D86FBB>
28. Fuente de imagen de estudio de variación SNP por SNP :  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210590/>
29. Fuente de imagen y página oficial del Proyecto 1000 Genoma :  
<http://www.internationalgenome.org/>
30. Información relacionada a Proyecto HapMap : <https://es.wikipedia.org/wiki/HapMap>
31. Información relacionada a Genoma Humano :  
[https://es.wikipedia.org/wiki/Genoma\\_humano](https://es.wikipedia.org/wiki/Genoma_humano)
32. Información relacionada a ácido desoxirribonucleico :  
[https://es.wikipedia.org/wiki/%C3%81cido\\_desoxirribonucleico](https://es.wikipedia.org/wiki/%C3%81cido_desoxirribonucleico)
33. Información relacionada a haplotipo y fuente de imagen de haplotipo:  
<https://www.genome.gov/27562906/acerca-del-proyecto-internacional-hapmap/>
34. Información relacionada a secuenciación genómica :  
<http://www2.iib.uam.es/seq/tecnicas/biomed1.html>
35. Información relacionada a medicina de precisión :  
<https://kidshealth.org/es/parents/precision-medicine-esp.html>
36. Información relacionada a trombosis venosa profunda en la biblioteca Nacional de Medicina de los Estados Unidos :  
<https://medlineplus.gov/spanish/ency/article/000156.htm>
37. Información relacionada a trombosis venosa profunda :  
<https://www.mayoclinic.org/es-es/diseases-conditions/deep-vein-thrombosis/symptoms-causes/syc-20352557>
38. Información relacionada a Ensembl : <https://www.ensembl.org/info/about/index.html>
39. Información relacionada a Ensembl :  
[https://en.wikipedia.org/wiki/Ensembl\\_genome\\_database\\_project](https://en.wikipedia.org/wiki/Ensembl_genome_database_project)
40. Información relacionada a ClinVar : <https://www.ncbi.nlm.nih.gov/clinvar/intro/>
41. Datos Estadísticos ClinVar: <https://www.ncbi.nlm.nih.gov/clinvar/submitters/>
42. Información relacionada a SNPedia :  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245045/>
43. GWAS Catalog : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210590/>
44. Información relacionada a Europe PMC : <http://europepmc.org/About>

45. Información relacionada a Proyecto Genoma Humano :  
<https://www.monografias.com/trabajos89/proyecto-genoma-humano-discriminacion/proyecto-genoma-humano-discriminacion.shtml>
46. Información relacionada a proyecto ENCODE :  
<https://www.encodeproject.org/about/contributors/>
47. Información relacionada a Genes : <http://www.dciencia.es/adn-genes-cromosomas/>
48. Información relacionada a Principio de Hardy-Weinberg :  
[https://es.wikipedia.org/wiki/Ley\\_de\\_Hardy-Weinberg](https://es.wikipedia.org/wiki/Ley_de_Hardy-Weinberg)
49. Información relacionada a Haplotipo : <https://es.wikipedia.org/wiki/Haplotipo>
50. Información relacionada a Odds Ratio : [https://en.wikipedia.org/wiki/Odds\\_ratio](https://en.wikipedia.org/wiki/Odds_ratio)
51. Información relacionada a Odds Ratio e interval de confianza:  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2545775/?page=1>
52. Página oficial del repositorio PubMed : <https://www.ncbi.nlm.nih.gov/pubmed/>
53. Información relacionada a DbSNP : <https://www.ncbi.nlm.nih.gov/projects/SNP/>
54. Fuente de imagen proceso de la terapia génica :  
<https://terapiagenica.science/proceso-de-la-terapia-genica/>
55. Fuente de imagen diagnóstico preimplantacional:  
<https://www.reproduccionasistida.org/procedimiento-del-dgp/>
56. Fuente de imagen técnica de secuenciación del genoma humano :  
<http://webs.ucm.es/info/genetica/AVG/practicas/secuencia/Secuencia.htm>
57. Fuente de captura de pantalla de datos genéticos pertenecientes al proyecto Hapmap publicados en la página oficial de Ensembl:  
[https://www.ensembl.org/Homo\\_sapiens/Variation/Population?db=core;r=10:112998090-112999090;y=rs7903146;vdb=variation;vf=4835099](https://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=10:112998090-112999090;y=rs7903146;vdb=variation;vf=4835099)
58. Fuente de imagen de medicina de precisión:  
<https://hitconsultant.net/2016/02/22/31535/>
59. Fuente de imagen de Adn desde la página oficial de instituto nacional de cáncer NIC :  
<https://www.cancer.gov/espanol/publicaciones/diccionario/def/histona>
60. Fuente de imagen de cromosoma humano:  
<http://sciencedoing.blogspot.com/2013/01/human-male-degenerative-sex.html>
61. Fuente de imagen con la definición gráfica del gen: <https://es.wikipedia.org/wiki/Gen>
62. Fuente de imagen con el resumen del proceso de traducción y transcripción:  
<http://www.monsantoglobal.com/global/py/productos/pages/transcripcion.aspx>
63. Fuente de imagen de alelos: <http://biologiabi.blogspot.com/2011/06/14-cromosomas-genes-alelos-y-mutaciones.html>
64. Fuente de imagen con algunas mutaciones cromosómicas:  
<http://benitobios.blogspot.com/2008/11/tipos-de-mutaciones.html>
65. Fuente imagen de genotipo a fenotipo:  
<http://marinaaboitiz.blogspot.com/2017/03/genotipo-y-fenotipo.html>
66. Fuente imagen de los síntomas de diabetes mellitus tipo 2: <http://tlvz.com/los-primeros-sintomas-de-advertencia-de-la-diabetes-tipo-1-y-tipo-2/>
67. Fuente imagen de Trombosis venosa profunda:  
<https://saludybienestareblog.com/2017/11/27/caso-1-la-medicacion-le-provoca-una-trombosis-venosa-profunda/>

68. Fuente imagen de estadística de Trombosis venosa profunda:  
<https://www.slideshare.net/GamalAgmy/updates-in-venous-thromboembolism>
69. Fuente de imagen y la página oficial de Ensembl: <https://www.ensembl.org/index.html>
70. Fuente de imagen de las regiones de interés en la página oficial de Ensembl:  
[https://www.ensembl.org/Homo\\_sapiens/Variation/Explore?db=core;r=10:112998090-112999090;v=rs7903146;vdb=variation;vf=4835099](https://www.ensembl.org/Homo_sapiens/Variation/Explore?db=core;r=10:112998090-112999090;v=rs7903146;vdb=variation;vf=4835099)
71. Fuente de imagen de la página oficial de ClinVar:  
<https://www.ncbi.nlm.nih.gov/clinvar/>
72. Fuente imagen de participantes en ClinVar :  
<https://www.ncbi.nlm.nih.gov/clinvar/docs/map/>
73. Fuente imagen página oficial de SNPedia:  
<https://www.ncbi.nlm.nih.gov/clinvar/docs/map/>
74. Fuente imagen página oficial de GWAS Catalog: <https://www.ebi.ac.uk/gwas/search>
75. Fuente imagen de enfermedad relacionada en Ensembl:  
[https://www.ensembl.org/Homo\\_sapiens/Variation/Explore?r=4:147586898-147587898;v=rs1040196;vdb=variation;vf=753153](https://www.ensembl.org/Homo_sapiens/Variation/Explore?r=4:147586898-147587898;v=rs1040196;vdb=variation;vf=753153)
76. Fuente imagen artículo de enfermedades relacionadas en GWAS Catalog:  
<https://www.ebi.ac.uk/gwas/search?query=rs10401969>
77. Fuente de imagen de enfermedad relacionada en PubMed:  
<https://www.ncbi.nlm.nih.gov/pubmed/?term=29221444>
78. Fuente imagen de SNP localizado fuera de gen en Ensembl:  
[https://www.ensembl.org/Homo\\_sapiens/Variation/Phenotype?db=core;r=11:2182558-2183558;v=rs11043007;vdb=variation;vf=6377354](https://www.ensembl.org/Homo_sapiens/Variation/Phenotype?db=core;r=11:2182558-2183558;v=rs11043007;vdb=variation;vf=6377354)
79. Fuente de imagen de SNP localizado en dbSNP: <https://www.ncbi.nlm.nih.gov/snp>
80. Fuente de imagen de localización del Gen en dbSNP:  
<https://www.ncbi.nlm.nih.gov/gene/?term=TH>
81. Fuente imagen de estudio SNP por SNP en GWAS Catalog :  
<https://www.ebi.ac.uk/gwas/search?query=23509962>
82. Fuente imagen del artículo publicado en Pubmed:  
<https://www.ncbi.nlm.nih.gov/pubmed/?term=25755232%5Buid%5D>
83. Fuente de imagen SNPedia con los valores estadísticos faltantes:  
<https://www.snpedia.com/index.php/Rs2028299>
84. Fuente imagen de variación con valores estadísticos distintos en GWAS Catalog:  
<https://www.ebi.ac.uk/gwas/search?query=rs1470579>
85. Fuente imagen Ensembl con valores estadísticos distintos:  
[https://www.ensembl.org/Homo\\_sapiens/Variation/Phenotype?db=core;r=3:185810792-185811792;v=rs1470579;vdb=variation;vf=1029813](https://www.ensembl.org/Homo_sapiens/Variation/Phenotype?db=core;r=3:185810792-185811792;v=rs1470579;vdb=variation;vf=1029813)

## ANEXO 1. PARÁMETROS FINALES PARA LA CARGA DE DATOS

	<b>Descripción</b>	<b>Ejemplo</b>
<b>DBSNP</b>	Identificador único de entrega proporcionado por dbSNP	rs18726354
<b>NC_IDENTIFIER</b>	Versión de la secuencia cromosómica respecto a la variación dada proporcionado por NCBI	NC_000002.11
<b>CHROMOSOME</b>	Nombre o número del cromosoma donde la variación esta mapeada.	1
<b>GRCh37_START</b>	Localización de inicio en el genoma de referencia GRCh37	6290728
<b>GRCh37_END</b>	Localización de fin en el genoma de referencia GRCh37	6290728
<b>SYNONYM</b>	Posibles sinónimos del gen afectado	FLJ21877,KIAA1767,GITA,ARMC13
<b>GRCh38_START</b>	Localización de inicio en el genoma de referencia GRCh38	22134095
<b>GRCh38_END</b>	Localización de fin en el genoma de referencia GRCh38	22134096
<b>RISK_ALLELE</b>	Para los rasgos de la enfermedad, definimos los alelos de riesgo como variantes que corresponden a un OR de enfermedad mayor que uno.	A (Adenina)
<b>NG_IDENTIFIER</b>	La secuencia del gen (así como su versión) proporcionado por NCBI	NG_052991.1
<b>NM_IDENTIFIER</b>	La secuencia del transcrito (así como su versión) proporcionado por NCBI	NM_001105250.2,NM_001272020.1,NM_001330195.1,NM_004796.5,NM_138970.4
<b>NP_IDENTIFIER</b>	La secuencia de la proteína proporcionado por NCBI	NP_660310.2
<b>REF</b>	Alelo de referencia	T
<b>ALT</b>	Alelo alternativo	C
<b>FREQ.GENO.1</b>	Es el alelo heredado en una posición del SNP en la fase 1	A A: 0.628
<b>VARIATION_TYPE</b>	Tipo de variación	SNP
<b>FREQ.GENO.2</b>	Es el alelo heredado en una posición del SNP en la fase 2	A G: 0.324
<b>FREQ.ALT</b>	Frecuencia de aparición del alelo alternativo perteneciente	G: 0.210
<b>FREQ.GENO.3</b>	Es el alelo heredado en una posición del SNP en la fase 3	G G: 0.048
<b>FREQ.REF</b>	Frecuencia de aparición del alelo de referencia perteneciente	A: 0.790
<b>PARTICIPANTS</b>	Tamaño inicial de sujetos en el estudio	159208

<b>POPULATION</b>	Origen geográfico de los sujetos en el estudio	<i>European</i>
<b>REPLICATED</b>	Si se ha realizado alguna replicación del estudio	YES
<b>P-VALUE</b>	Nivel de evidencia de la relación genotipo-fenotipo asignada por el estudio	2,00E-06
<b>ORIGIN</b>	Origen de la mutación	GERMLINE
<b>OR</b>	Es la medida utilizada para indicar si la presencia o falta de algo está relacionada con la presencia o falta de otro factor [23]	1.1
<b>CI</b>	Es el rango de confianza	[0.016-0.162]
<b>PUBMED_ID</b>	Identificador de la referencia bibliográfica en la base de datos PubMed	29221444
<b>YEAR</b>	Año de publicación	2017
<b>PUBMED_URL</b>	Enlace al artículo o estudio por medio del identificados	<a href="http://www.ncbi.nlm.nih.gov/pubmed/28566273">www.ncbi.nlm.nih.gov/pubmed/28566273</a>
<b>REFERENCE</b>	Autores y entidades participantes en la investigación	Sim X, Ong RT, Suo C, Tay WT, Liu J, Ng DP, Boehnke M, Chia KS, Wong TY, Seielstad M, Teo YY, Tai ES.
<b>GENE</b>	Bases de datos de las que se extrae al información	ClinVar
<b>GENE_DATABASE</b>	Base de datos en la que se encuentra identificado el gen.	HGNC
<b>GEN</b>	Siglas utilizadas para referenciarse al gen mapeado donde se encuentra la variación	NRXN3
<b>GEN_NAME</b>	Nombre completo del gen mapeado donde se encuentra la variación	neurexin 3
<b>GEN_ID</b>	Identificador único asociado al gen por HGNC	HGNC:8010
<b>PHENOTYPE</b>	Enfermedades a las que se asocia	TYPE 2 DIABETES
<b>SIGNIFICADO CLÍNICO</b>	El efecto clínico de la variación	<i>Pathogenic</i>
<b>CREATION_DATE</b>	Fecha de creación de la variación	05/12/2003
<b>UPDATE_DATE</b>	Última fecha de modificación de la variación	23/01/2018
<b>GWAS_ID</b>	Identificador asociado al artículo o estudio en el repositorio GWAS	GCST004773

TABLA 7. PARÁMETROS FINALES PARA LA CARGA DE DATOS



## ANEXO 2. PREPARACIÓN DE DATOS PARA LA CARGA

Algunos de los campos que faltaban para la carga de datos fueron incorporados de la siguiente forma:

- ❖ En caso de **alelo de referencia** y el **alelo alternativo** se consulto el repositorio Ensembl y en la sección principal de la variación se indican los dos alelos separados por una barra como se puede ver en la Ilustración 41.

Variant: rs3843467

**rs3843467** SNP

Most severe consequence [intron variant](#) | [See all predicted consequences](#)

Alleles **G/T** | Ancestral: T | MAF: 0.20 (T) | Highest population MAF: 0.33

Location [Chromosome 5:56560548](#) (forward strand) | VCF: 5 56560548 rs3843467 G T

Evidence status

HGVS names This variant has 3 HGVS names - [Hide](#)

Ensembl HGVS: [NC\\_000005.10:g.56560548G>T](#) [ENST00000438651.5:c.-15-4361C>A](#)

dbSNP HGVS: [NM\\_001287053.1:c.-15-4361C>A](#)

ILUSTRACIÓN 41. ALELO DE REFERENCIA Y ALELO ALTERNATIVO EN ENSEMBL

- ❖ Los datos relacionados con las **frecuencias de los alelos** y las **frecuencias genotípicas** estaban en el apartado de 'Population Genetics' del repositorio Ensembl. En este caso estaban presente las estadísticas de los estudios como Proyecto 1000 genomas (ilustración 42), HapMap, TopMed u otros. Pero en este trabajo siempre se ha intentado referenciar las estadísticas del Proyecto 1000 Genoma y en caso de no disponer de esta se optó por los datos proporcionados por el proyecto HapMap.

Population	Allele: frequency (count)	Genotype: frequency (count)
ALL	G: 0.799 (4003) T: 0.201 (1005)	G G: 0.640 (1603) G T: 0.318 (797) T T: 0.042 (104)
AFR	G: 0.765 (1011) T: 0.235 (311)	G G: 0.576 (381) G T: 0.377 (249) T T: 0.047 (31)
ACB	G: 0.740 (142) T: 0.260 (50)	G G: 0.542 (52) G T: 0.396 (38) T T: 0.062 (6)
ASW	G: 0.779 (95) T: 0.221 (27)	G G: 0.590 (36) G T: 0.377 (23) T T: 0.033 (2)
ESN	G: 0.758 (150) T: 0.242 (48)	G G: 0.596 (59) G T: 0.323 (32) T T: 0.081 (8)
GWD	G: 0.748 (169) T: 0.252 (57)	G G: 0.540 (61) G T: 0.416 (47) T T: 0.044 (5)
LWK	G: 0.788 (156) T: 0.212 (42)	G G: 0.606 (60) G T: 0.364 (36) T T: 0.030 (3)
MSL	G: 0.800 (136) T: 0.200 (34)	G G: 0.612 (52) G T: 0.376 (32) T T: 0.012 (1)
YRI	G: 0.755 (163) T: 0.245 (53)	G G: 0.565 (61) G T: 0.380 (41) T T: 0.056 (6)
AMR	G: 0.716 (497) T: 0.284 (197)	G G: 0.516 (179) G T: 0.401 (139) T T: 0.084 (29)

ILUSTRACIÓN 42. FRECUENCIA DE REFERENCIA, FRECUENCIA DEL GEN ALTERNATIVO, FRECUENCIA GENOTIPO 1, FRECUENCIA GENOTIPO 2 Y FRECUENCIA GENOTIPO 3 EN ENSEMBL


- ❖ El nombre perteneciente a la variante del repositorio dbSNP proporcionado por HGVS se encontraba en la página de Ensembl como se puede ver en la ilustración 43.

### rs3843467 SNP

Most severe consequence | [intron variant](#) | [See all predicted consequences](#)

Alleles | [G/T](#) | Ancestral: T | MAF: 0.20 (T) | Highest population MAF: 0.33

Location | [Chromosome 5:56560548](#) (forward strand) | VCF: 5 56560548 rs3843467 G T

Evidence status ⓘ |  gnomAD

HGVS names | This variant has 3 HGVS names - [Hide](#) ☒

Ensembl HGVS: | dbSNP HGVS:

- [NC\\_000005.10:g.56560548G>T](#)
- [ENST00000438651.5:c.-15-4361C>A](#)
- [NM\\_001287053.1:c.-15-4361C>A](#)

ILUSTRACIÓN 43. NOMBRE DE LA VARIANTE EN ENSEMBL

- ❖ La referencia a los artículos y estudios es otro de los campos que se deben proporcionar para la carga de datos y en este caso se localizaron casi todas en apartado de Citaciones de Ensembl indicado en la ilustración 44.

### Citations

rs3843467 is mentioned in the following publications

Year	PMID	Title	Author(s)	Full text
2018	<a href="#">29038864</a>	Improved detection of genetic loci in estimated glomerular filtration rate and type 2 diabetes using a pleiotropic cFDR method.	<a href="#">Liu HM, He JY, Zhang Q, Lv WQ, Xia X, Sun CQ, Zhang WD, Deng HW.</a>	<a href="#">PMC5819009</a>

ILUSTRACIÓN 44. REFERENCIAS DEL AUTORES EN ENSEMBL

## ANEXO 3. RESUMEN PROCESO IDENTIFICACIÓN

Fase	Iniciales	Finales	Eliminadas
<b>Criterios de calidad (5.3.1.1)</b>	27581	21029	6552
<b>Enfermedades Relacionadas (5.3.2.1)</b>	21029	18448	2581
<b>Afectados por proximidad (5.3.2.2)</b>	18448	17447	1001
<b>Estudio combinación de variaciones (5.3.2.3)</b>	17447	17447	0
<b>Citaciones (5.3.2.6)</b>	17447	10372	7075
<b>Población y fases (5.3.2.7)</b>	10372	8435	1937
<b>Criterios de Calidad (5.3.3)</b>	8435	1103	7332
<b>Merge (5.4)</b>	1103	1093	10
<b>Finales</b>	1093		

TABLA 8. PROCESO IDENTIFICACIÓN DIABETES MELLITUS TIPO 2

Fase	Iniciales	Finales	Eliminadas
<b>Criterios de calidad (5.3.1)</b>	1247	1188	59
<b>Enfermedades Relacionadas (5.3.2.1)</b>	1188	1188	0
<b>Afectados por proximidad (5.3.2.2)</b>	1188	1188	0
<b>Estudio combinación de variaciones (5.3.2.3)</b>	1188	1058	130
<b>Citaciones (5.3.3.6)</b>	1058	649	409
<b>Población y fases (5.3.2.7)</b>	649	497	152
<b>Criterios de Calidad (5.3.3)</b>	497	155	342
<b>Merge (5.4)</b>	155	155	0
<b>Finales</b>	155	-	-

TABLA 9. PROCESO IDENTIFICACIÓN TROMBOSIS VENOSA PROFUNDA