DISEÑO Y DESARROLLO DE UN SISTEMA DE INFORMACIÓN GENÓMICA BASADO EN UN MODELO CONCEPTUAL HOLÍSTICO DEL GENOMA HUMANO

POR: JOSÉ FABIÁN REYES ROMÁN

PROS
Centro de Investigación en Métodos de Producción de Software

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Supervisor:
Óscar Pastor López

Febrero 2018

Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano

José Fabián Reyes Román

TESIS DOCTORAL DEPOSITADA EN CUMPLIMIENTO PARCIAL DE LOS REQUISITOS PARA EL GRADO DE DOCTOR EN INFORMÁTICA

Dirigida por:
Prof. Dr. Óscar Pastor López
opastor@dsic.upv.es

Febrero 2018

Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano

Design and Development of a Genomic Information System Based on a Holistic Conceptual Model of the Human Genome

Disseny i Desenrotllament d'un Sistema d'Informació Genòmica Basat en un Model Conceptual Holístic del Genoma Humà

Tesis defendida por José Fabián Reyes Román el 12 de febrero del 2018 para la obtención del título de Doctor en Informática por la Universitat Politècnica de València.

Supervisor
Prof. Dr. Óscar Pastor López, Universitat Politècnica de València, España

Comité de Evaluadores Externos:
Prof. Dr. Manuel Noguera, Universidad de Granada, España
Prof. Dr. Manuel Pérez Alonso, Universitat de València, España
Prof. Dr. Juan Carlos Trujillo Mondéjar, Universitat d'Alacant, España

Miembros del Tribunal de la Tesis:
Presidente: Prof. Dr. José Ignacio Panach N., Universitat de València, España
Secretario: Prof. Dr. Juan Carlos Trujillo Mondéjar, Universitat d'Alacant, España
Vocal: Prof. Dr. José Luís Garrido Bullejos, Universidad de Granada, España

Suplentes – Evaluadores Externos & Tribunal de Defensa:
Prof. Dr. Jesús Peral Cortés, Universitat d'Alacant, España
Prof. Dr. Johann Eder, Alpen Adria Universität Klagenfurt, Austria
Prof. Dra. Ruth Cobos Pérez, Universidad Autónoma de Madrid, España
Prof. Dra. Tanja Ernestina Vos, Universitat Politècnica de València, España

PROS
Centro de Investigación en Métodos de Producción de Software

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

Este informe fue preparado por:

José Fabián Reyes Román
jreyes@pros.upv.es
Centro de Investigación en Métodos de Producción de Software (PROS)
Departamento de Sistemas Informáticos y Computación (DSIC)
Universitat Politècnica de València
Camino de Vera s/n, Edificio 1F, DSIC
46022, Valencia, España

Diseño de portada:
José F. Reyes Román, José Wilmar Vera V. & Dinamika Estudios

Implementación por:
Dinamika Estudios (*dinamo.director@gmail.com*)

Comentarios:

- Esta tesis doctoral ha sido financiada por el *Ministerio de Educación Superior, Ciencia y Tecnología* (MESCyT) de Santo Domingo, República Dominicana.

"Porque para Dios no hay nada imposible."
Lucas 1:37

"Y sabemos que a los que aman a Dios, todas las cosas les ayudan a bien, esto es, a los que conforme a su propósito son llamados."
Romanos 8:28

A Dios
A mi familia
A ti mi reina, mi amada esposa…

Prefacio

Esta tesis doctoral es el resultado de muchos **sueños** (*presentes*) a lo largo de mi vida. Es la combinación de muchos ingredientes como: la **fe**, **perseverancia**, **constancia**, deseos de **superación** y **crecimiento**, no sólo en lo profesional sino también en lo personal. Definitivamente, es la materialización de largas horas de **esfuerzo**, **dedicación** y **sacrificio** en mi aventura como investigador *-algo que estaba escrito en los planes de Dios-*; es el resultado palpable de las muestras de amor y palabras de ánimo que fueron lanzadas sobre mi vida en el momento preciso (exacto) por grandes personas *-valiosas* e *inolvidables-*.

Al escribir estas líneas se llena de **alegría** y **emoción** mi corazón, pues me transporta a mis 8 años (1997) cuando recién iniciaba mi primer curso de verano para aprender a utilizar el Windows 95, allí fue donde realmente comenzó mi **pasión** por todo lo relacionado con la **Informática**. De esta forma continué buscando cursos y talleres sobre *paquetes ofimáticos*, *reparación y mantenimiento de PC*, *entre otros* que me ayudaron a entender mejor cómo funcionaba el *mundo de los ordenadores*. Mientras finalizaba mi primer ciclo de estudios de secundaria me enteré a través de unos buenos amigos de la posibilidad de concluir el bachillerato con mención de "*técnico en informática*", por lo que no dude en cambiarme de escuela y lanzarme a esa nueva aventura. Fueron dos años (2004-2006) espectaculares, en donde aprendí desde el funcionamiento de los ordenadores (*hardware*) hasta la parte lógica de la programación (*software*).

Hasta aquí parecía estar todo claro, pero no fue así… Al acceder a la universidad me sentía atraído por las áreas de la *salud*, pero no me sentía seguro sobre embarcar un nuevo viaje en un área totalmente distinta. Por lo que finalmente realicé mi matrícula en el grado de *Ingeniería de Sistemas* (UCE), en donde si sentía que era mi lugar.

Ahora bien, a lo mejor te preguntas: *¿Y por qué todo este preámbulo?* Bueno, lo comparto porque es que me ha quedado completamente claro que definitivamente los *planes de* ***Dios*** son de bien y no de mal

como nos dice su palabra (*Jeremías 29:11*), y finalizar mis estudios de doctorado me demuestran que el Señor tiene grandes sueños para cada uno de nosotros y que con su ayuda podemos alcanzarlos. Después de mi carrera decidí dar un paso más en el escalón de la *informática*, mediante la realización de mi máster en la UPV. Un máster que me dio la oportunidad de conocer a mi director (Julio 2012), un excelente investigador y ser humano, el cual no dudó en ningún momento el darme la oportunidad de iniciar una carrera investigadora en la cual yo pudiera aportar mi ***granito de arena*** al mundo científico.

De la mano de mi director y compañeros de grupo (*genoma*) hemos aprendido a utilizar nuestros conocimientos para aplicarlos al área de la *salud* -en este punto encontré la **conexión** de mi **pasión** por la **informática** y como podría explotarla para contribuir paralelamente al ámbito de la **salud**-. Por lo que trabajar en el **Diseño y Desarrollo de Sistemas de Información Genómicos** aplicando técnicas de **Modelado Conceptual** ha sido para mí una experiencia muy gratificante y enriquecedora. Ha sido una aventura en donde he trabajado y aprendido mucho sobre cómo entender nuestro **genoma** y que a través del uso de las últimas tecnologías en *sistemas de información* e *ingeniería de software* se puede repercutir de manera positiva en la calidad de vida de los seres humanos.

Este trabajo no se detiene aquí, cada día hay nuevas metas y desafíos muy interesantes… El dominio genómico es bastante *amplio* y *complejo*, pero gracias al trabajo hecho hasta el día de hoy y el que se continúa desarrollando podemos ver a la vuelta de la esquina la tan aclamada "*Medicina de Precisión*".

"*El que no vive para servir, no sirve para vivir.*"
Teresa de Calcuta
Marcos 10:45

Agradecimientos

En el camino de la vida nos encontramos con muchas personas que nos ayudan a *aprender*, *crecer* y *avanzar*, y en el trayecto de esta aventura han sido muchos los momentos especiales y únicos compartidos con cada uno ustedes.

En primer lugar, doy las gracias a *Dios*, por estar conmigo en todo momento, por ser mi fortaleza, mi escudo, mi refugio, mi padre y amigo. Porque sin él nada de esto hubiera sido posible. Hoy puedo ver y disfrutar de un sueño que un día pusiste en mi corazón, el cual he alcanzado gracias a tu compañía, ayuda, ánimo, fuerzas… Porque he sentido tu amor en los momentos en que más lo he necesitado. Gracias por no dejarme solo ni un instante de mi vida. (*Isaías 41:13*)

A Mis padres *José Fabián Reyes Olea & Nilda Román Cedeño*, gracias por soñar conmigo, ¡*lo hemos logrado*! Gracias por ser un apoyo incondicional en toda mi vida. Son los mejores padres y los amo con todo mi corazón. Gracias por ser un ejemplo para mí, por enseñarme que, con esfuerzo, dedicación y sacrificio los sueños se pueden alcanzar. Este *triunfo* es de ustedes… Sus oraciones y palabras de ánimo han sido fundamental para llegar a este día. Gracias por enseñarme buenos principios y por instruirme por el camino del bien. (*Colosenses 3:20*)

A mi amada esposa *Claudia M. Agudelo Sterling*, ¡*mi amor lo logramos*! Esta victoria es tuya… Eres mi mayor bendición, mi amiga, mi compañera de batallas, mi tesoro… Definitivamente he llegado aquí porque Dios te envió en el momento justo, él sabía que contigo al lado llegaría más lejos. Gracias por cada tiempo vivido, gracias por tus palabras (consejos) y por inyectarme de Fe cuando me faltaba. Te amo con todas mis fuerzas, y te agradezco por estar a mi lado en todo momento. "*Eres la mejor*". (*Eclesiastés 4:11-13*)

A mi hermana *Nilda Josefina Reyes Román "Vanessa"*, gracias por estar allí y por ayudarme siempre que has podido. Gracias por cuidar y apoyar a nuestros padres, por acompañarlos en momentos en los cuales no he podido estar… ¡*Gracias por todo*! (*Josué 1:6-9*)

A mi familia *Reyes Olea & Román Cedeño*, gracias por formar parte de todo lo que soy, gracias por cada mensaje o llamada. Este logro lo quiero dedicar de forma muy especial a mis abuelas *Clara Lucía Olea & Genara Cedeño*, hoy no están físicamente con nosotros, pero siguen viviendo en nuestros corazones -*Mi manita y mamá*-, gracias por ser tan especiales y amorosas conmigo, por todas las tardes de aventuras, conversaciones y risas compartidas. También quiero agradecer de forma especial a todos mis tíos (as) y primos, porque siempre estuvieron allí para escucharme, y ayudarme a seguir adelante a pesar de las distancias. Gracias a ti *Odeydris Adeyanira Ortíz Reyes*, por estar pendiente de mi durante toda mi estancia en España, muchas gracias por tu cariño y por tus palabras.

A mi familia *Agudelo Sterling*, gracias por acogerme como parte de la familia, por cada momento compartido y por ser parte de este sueño. Gracias por su apoyo y cariño. *Guillermo & Orfa* darles las gracias por sus consejos y sus muestras de amor, a mis cuñis *Diana & Lina*, gracias por todos los momentos compartidos y por su amistad, gracias por compartir con nosotros las vivencias y aventuras de nuestro sobrino *Emanuel*.

A mi director *Dr. Óscar Pastor López*, gracias por todo. Gracias por darme el privilegio de aprender tanto de usted y poder formar parte de su grupo de investigación. Son incalculables las enseñanzas y experiencias vividas, muchas veces veía este momento inalcanzable, pero siempre tuvo Fe en que lo lograríamos. ¡Hemos llegado! Recuerdo ese julio 2012 donde emprendía un nuevo camino en el dominio genómico (en ese entonces sólo como una tesis de máster), hasta hoy que vemos materializado este sueño, el cual ha sido alcanzado gracias a su confianza y apoyo incondicional. Gracias por ayudarme a crecer como profesional e investigador, y por ser junto a *Carmen* parte de mi gran familia en València.

A mi familia espiritual *Iglesia Rey de Reyes de Valencia*, gracias amados pastores *Antonio & Titi Selma*, gracias por ser unos padres para mí, por brindarme tantas muestras de amor y cariño. Gracias por cada palabra de parte de Dios, han sido refrigerio y renuevo para mí. Gracias por ayudarme y permitirme crecer junto a ustedes. Gracias a

todos los líderes y hermanos de la iglesia por cada minuto compartido, gracias por cada abrazo, por las risas y muestras de afecto, en RDR somos una gran familia. *Carlos Hugo Cuellar & Jenny Paz*, *Aymer Zabala & Lucía Sánchez* y *Óscar H. Tenorio & Mónica J. Calvo* gracias por sus consejos, por ser un apoyo en momentos claves de mi vida y matrimonio. *Nahúm Peña & Judennis Olivo*, gracias por su amistad y por estar al pendiente de nosotros. ¡Dios les bendiga!
También quiero agradecer a los hermanos y compañeros de ministerio, en especial a mis líderes *Carlos E. Paz, Bertha Añez & Ana Paola Gutiérrez*, gracias por su apoyo y por entenderme en situaciones en las cuales no podía estar. Gracias a mis compañeros del 1º y 2º Equipo del Pastor por todas las enseñanzas y lecciones aprendidas. A los grupos J2 & H2, gracias chicos por formar parte de este sueño, por sus oraciones y palabras de ánimo, quiero agradecerles de forma especial al equipo: *Yannick Agoli-Agbo*, *Ricardo Enguídanos Creo* & *Willian E. Martínez*.

A mis amigos y hermanos *Robert Adolfo Santana R., Jhordyn Miguel Contreras J., Wilson Germán Ramírez, Melvin U. Vargas Sánchez & In4+Tik'06*, gracias por ser unos hermanos para mí, grandes amigos que han estado en las buenas y en las no tan buenas… Gracias por su apoyo incondicional, por las largas horas hablando y riendo. Gracias por ser genuinos y transparentes.

A *Fausto I. Nelson A. & Alexander Santana Donato*, gracias por todas las experiencias vividas, junto a ustedes empecé este viaje (2011) y agradezco el hecho de que a pesar de la distancia siempre estuvieran pendientes de los avances que iba alcanzando en este camino doctoral. Gracias por sus oraciones. También quiero agradecer a nuestros compañeros del máster (MISMFSI) por ser un grupo único y unido, en el cual se crearon grandes lazos de amistad *-a nivel internacional-*, gracias especialmente a mis amigos y hermanos *Eugenio Bolaños Cárdenas* (Colombia) & *José Alberto Martí Martín* (València).

A mi familia en *Valencia*, el estar lejos de tus seres queridos es muy complicado, pero gracias a Dios que coloca personas que te aman y se preocupan por ti, grandes amigos que nos ayudan a que nuestro caminar sea mucho más llevadero. Es por ello, que quiero agradecer a esas personas que durante estos años se han convertido en parte de mi familia… A *Willian E. Martínez V. & Erica Del Pup*, gracias *xiquets*

por todos los momentos y risas compartidos, gracias por su amistad y por el apoyo que nos han brindado en todo tiempo a Molz y a mí. Willy gracias por ser un gran amigo y hermano en esta tierra. *Rosmeyris Vargas, nuestro ahijado Aarón & Familia*, gracias por ser tan especiales con nosotros y por hacernos sentir parte de su familia. *Joel Pérez & familia*, gracias por tus consejos y por contar siempre conmigo. *Mariana Rivero & Familia*, gracias por estos años de amistad y por todas las vivencias compartidas. A mis hermanitas *Vanessa Buitrago (Youuu Youuu), Jelissa Montero y Familiares*, porque desde que llegue hemos compartido y crecido juntos, gracias por sus palabras y por compartir juntos este sueño. *Olga Dipre & Jonathan Cuevas*, gracias por su cariño y por contar conmigo. ¡Dios les bendiga y recompense!
También quiero agradecer a todos aquellos que de una forma u otra han creado ese ambiente familiar para disfrutar de un pedacito de República Dominicana en esta hermosa nación (*Keyla*, *Billceydi*, *Massiel*, *Loanny S.*, *Elizabeth*, *Mamá Ramona* y todos aquellos que ya han retornado).

Al *Centro de Investigación en Métodos de Producción de Software (PROS)*, gracias por el apoyo incondicional durante el desarrollo de esta tesis doctoral. Gracias a *Ana Cidad Vila*, por su colaboración y ayuda durante todos estos años, por gestionar todo en el centro de manera eficiente y productiva.

A *mis compañer@s y amig@s del PROS*, gracias a todos aquellos que estuvieron desde mis inicios en *GemBiosoft*, cuando apenas iniciaba mis pasos en el mundo de la bioinformática, así como a todos mis compañeros del laboratorio 2L04 @DSIC. Quiero agradecerles de manera muy especial a cada uno de ustedes, porque definitivamente son parte esencial de este logro. Gracias por cada día compartido, por las comidas en *La Vella*, por escucharme, aconsejarme y sobre todo por sus palabras de ánimo. Gracias mil *Carlos I.*, *Ana L.*, *Lenin S.*, *Beatriz M.*, *Alberto G.*, *Urko R.*, *Julio S.*, *Ángel C.*, *Sipan*, *Ignacio²*, *Alexa*… ¡sois unos *cracks*! Doy las gracias a aquellos que estuvieron junto a nosotros @PROS pero que por circunstancias de la vida han emprendido el vuelo a otros lugares, gracias porque a través de ustedes he aprendido mucho, ver la perseverancia en cada uno de ustedes hasta lograr "*dar a luz*" a la tesis o algún proyecto de investigación fue de gran inspiración para mí, gracias *María Fernanda*, *Otto P.*, *Verónica*

B., *Faber*, *Marce*, *Sergio*, *MariaJo*. También quiero agradecer al equipo de profesores *Seniors*, *Juan Carlos Casamayor R.*, *Matilde Celma G. & Laura Mota*, gracias por las reuniones tan enriquecedoras sobre modelado conceptual y por su disposición para solucionar y discutir dudas del MCGH.

A *Francisco Valverde Girome & David Roldán Martínez*, gracias por sus revisiones, retroalimentación y colaboración en el desarrollo de esta tesis doctoral. Gracias por las reuniones y momentos compartidos.

A *José Marín Navarro @Luxemburgo*, gracias por su tiempo y por sus comentarios (revisiones) en esta tesis, los cuales han sido de gran valor para mí. Gracias mil por sus palabras tan alentadoras en momentos claves.

A *Mercedes R. Fernández A., José Francisco Santana V. & Kettsy C. Borrero*, gracias por sus aportes y enseñanzas a lo largo de mi vida profesional. Gracias por estar siempre al pendiente de mis estudios, y por cada mensaje alentador desde la distancia.

Al *Departamento de Sistemas Informáticos y Computación (DSIC)*, gracias a todo el personal por sus atenciones y por brindar todas las facilidades para trabajar de una manera más productiva.

A la *Oficina de Acción Internacional (OAI)-UPV*, gracias por el excelente trabajo realizado durante estos años, gracias por acogernos desde nuestra llegada a Valencia y permanecer junto a cada uno de nosotros (becarios) hasta la consecución de nuestra meta. Quiero dar las gracias de forma especial a *Geraldine Bustamante Reyes, Esther Durá Olcina & Ana Galiano Mainer*, por su disposición y colaboración en momentos esenciales de este sueño, gracias por escucharme cuando lo necesitaba y por alentarme a continuar hasta el final.

Al *Ministerio de Educación Superior, Ciencia y Tecnología (MESCyT) de la República Dominicana*, por la confianza puesta en mi para la realización del máster y posteriormente esta tesis doctoral. Gracias por la inversión realizada para la materialización de este logro.

A los *miembros del tribunal y del comité de evaluación*, gracias por haber aceptado participar en la fase final de mi tesis doctoral, ha sido un gran privilegio y honor contar con cada uno de ustedes.

Finalmente, quiero darles las gracias a todos ustedes que de una forma u otra han contribuido para que este sueño hoy fuera una realidad _____ (*pon tu nombre aquí*).

"El que no ama, no ha conocido a Dios; porque Dios es amor."
1 Juan 4:8

Resumen

Entender el genoma es un desafío de primer nivel, y esto se debe en gran parte a la gran cantidad de información existente en el dominio. Gracias a la aplicación de tecnologías NGS (*Next-Generation Sequencing*) se han generado enormes cantidades de datos -nuevos-, por lo que es fundamental construir estructuras que permitan *organizar*, *procesar* y *explorar* los datos con el fin de lograr un máximo provecho de la información y mejorar la comprensión del genoma humano.

En este estudio se define un marco de trabajo centrado en el uso del *Modelado Conceptual* como estrategia esencial para la búsqueda de soluciones. En el campo médico este enfoque de desarrollo de software está ganando impulso por su impacto en el trabajo realizado por *genetistas*, *laboratorios clínicos* y *bioinformáticos*.

Entender el genoma es un dominio de aplicación muy interesante debido a dos aspectos fundamentales: 1) en primer lugar, por las implicaciones sociológicas que supone plantearse la posibilidad de entender el lenguaje de la vida. 2) y, en segundo lugar, desde una perspectiva más práctica de aplicación en el ámbito clínico, debido a su repercusión en la generación de diagnósticos genómicos, los cuales juegan un papel importante dentro de la *Medicina de Precisión*.

En esta Tesis Doctoral se propone utilizar un *Modelo Conceptual del Genoma Humano* (MCGH) como base fundamental para la generación de *Sistemas de Información Genómicos* (GeIS), con el objetivo de facilitar una conceptualización del dominio que permita i) alcanzar un conocimiento preciso del dominio y ii) ser capaces de llegar a una medicina de precisión (personalizada). Es importante resaltar que este Modelo Conceptual debe permanecer en constante crecimiento debido a los nuevos aportes que surgen en la comunidad científica.

En este trabajo de investigación se presenta la evolución natural del modelo, así como un ejemplo de extensión del mismo, lo que permite comprobar su extensibilidad conservando su definición inicial. Además, se aplica el uso de una metodología (SILE) sistemática para la obtención de los datos desde los distintos repositorios genómicos, los cuales serán explotados a través herramientas software basadas en modelos conceptuales.

Mediante el uso de este *Modelo Conceptual holístico del Genoma Humano* se busca comprender y mejorar el compromiso ontológico con el dominio –*genómico*-, y desarrollar *Sistemas de Información Genómicos* apoyados en *Modelo Conceptuales* para ayudar a la toma de decisiones en el entorno bioinformático.

Resum

Entendre el genoma és un desafiament de primer nivell, i açò es deu en gran part a la gran quantitat d'informació existent en el domini. Gràcies a l'aplicació de tecnologies NGS (*Next-Generation Sequencing*) s'han generat enormes quantitats de dades - nous-, per la qual cosa és fonamental construir estructures que permeten *organitzar*, *processar* i *explorar* les dades a fi d'aconseguir un màxim profit de la informació i millorar la comprensió del genoma humà.

En este estudi es definix un marc de treball centrat en l'ús del *Modelatge Conceptual* com a estratègia essencial per a la busca de solucions. En el camp mèdic este enfocament de desenvolupament de programari està guanyant impuls pel seu impacte en el treball realitzat per *genetistes*, *laboratoris clínics* i *bioinformàtics*.

Entendre el genoma és un domini d'aplicació molt interessant a causa de dos aspectes fonamentals: 1) en primer lloc, per les implicacions sociològiques que suposa plantejar-se la possibilitat d'entendre el llenguatge de la vida. 2) i, en segon lloc, des d'una perspectiva més pràctica d'aplicació en l'àmbit clínic, a causa de la seua repercussió en la generació de diagnòstics genòmics, els quals juguen un paper important dins de la *Medicina de Precisió*.

En esta Tesi Doctoral es proposa utilitzar un *Model Conceptual del Genoma Humà* (MCGH) com a base fonamental per a la generació de *Sistemes d'Informació Genòmics* (GeIS), amb l'objectiu de facilitar una conceptualització del domini que permeta i) aconseguir un coneixement precís del domini i ii) ser capaços d'arribar a una medicina de precisió (personalitzada). És important ressaltar que este Model Conceptual ha de romandre en constant creixement degut a les noves aportacions que sorgixen en la comunitat científica.

En este treball d'investigació es presenta l'evolució natural del model, així com un exemple d'extensió del mateix, la qual cosa permet comprovar la seua extensibilitat conservant la seua definició inicial. A més, s'aplica l'ús d'una metodologia (SILE) sistemàtica per a l'obtenció de les dades des dels distints reposadors genòmics, els quals seran

explotats a través ferramentes de programari basades en models conceptuals.

Per mitjà de l'ús d'este *Model Conceptual holístic del Genoma Humà* es busca comprendre i millorar el compromís ontològic amb el domini -*genòmic*-, i desenvolupar *Sistemes d'Informació Genòmics* recolzats en *Model Conceptuals* per ajudar a la presa de decisions en l'entorn bioinformàtic.

Abstract

Understanding the genome is a first level challenge, and this is due in large part to a large amount of information in the domain. Thanks to the application of NGS (*Next-Generation Sequencing*) technologies, enormous amounts of -*new*- data have been generated, so it is essential to building structures that allow *organizing*, *processing* and *exploring* the data in order to obtain maximum benefit from the information and improve the understanding of the human genome.

In this study we define a framework focused on the use of *Conceptual Modeling* as an essential strategy for finding solutions. In the medical field, this approach to software development is gaining momentum due to its impact on the work carried out by *geneticists*, *clinical laboratories*, and *bioinformatics*.

Understanding the genome is a domain of very interesting application due to two fundamental aspects: 1) firstly, because of the sociological implications of considering the possibility of understanding the language of life. 2) secondly, from a more practical perspective of application in the clinical field, due to its repercussion in the generation of genomic diagnoses, which play an important role within *Precision Medicine*.

In this PhD, it is proposed to use a *Conceptual Model of the Human Genome* (CMHG) as the fundamental basis for the generation of *Genomic Information Systems* (GeIS), with the aim of facilitating a conceptualization of the domain that allows i) to achieve a precise knowledge of the domain and ii) be able to increase and improve the adaptation of genomics in personalized medicine. It is important to highlight that this Conceptual Model must remain in constant growth due to the new contributions that arise in the scientific community.

In this research work the natural evolution of the model is presented, as well as an example of its extension, which allows verifying its extensibility while preserving its initial definition. In addition, the use of a systematic methodology is applied to obtain the data from the different genomic repositories, which will be exploited through software tools based on conceptual models.

Through the use of this *Holistic Conceptual Model of the Human Genome*, we seek to understand and improve the ontological commitment to the -*genomic*- domain, and develop GeIS supported in *Conceptual Model* to help decision making in the bioinformatic environment in order to provide better treatment to the patients.

Tabla de Contenidos

Índice de figuras

Índice de tablas

CAPÍTULO 1

Motivación

Con el transcurso de los años se puede identificar la importancia que han tomado los *Sistemas de Información* (SI), lo cual se debe a su contribución directa en la mejora de procesos y/o gestión de los datos.

Un *sistema de información* se define como un sistema que *recopila*, *almacena*, *procesa* y *distribuye* información [1].

Gracias al gran número de ventajas que estos facilitan, se han implementado en una amplia diversidad de áreas de trabajo, como, por ejemplo, en el área de la salud, dando lugar a resultados muy positivos en la manipulación de los datos asociados al dominio de aplicación. Por esta razón, la comunidad científica y médica han unido sus esfuerzos con expertos en el tema con el objetivo principal de aprovechar estos conocimientos y aplicarlos al dominio "*genómico*".

El dominio genómico es un ejemplo palpable de la dificultad con la cual se enfrentan los especialistas al momento de gestionar las grandes cantidades de datos que se generan cada día. Por esta razón es esencial desarrollar sistemas de información que permitan explotar los datos de

una forma más efectiva. A raíz de todo el trabajo desarrollado en este contexto surgen los denominados *Sistemas de Información Genómicos* (GeIS, "*Genomic Information Systems*").

El uso de enfoques basados en *modelos conceptuales*[1] es una excelente estrategia para el desarrollo de sistemas potentes, ya que la representación conceptual de un dominio da lugar a un mayor entendimiento del mismo. Además, de que estos modelos permiten integrar -*de forma sencilla*- sobre la proyección inicial (ejemplo, versión 1 del modelo) los conocimientos generados por la evolución natural y constante de los datos genómicos.

El objetivo principal de esta tesis es facilitar un *Modelo Conceptual del Genoma Humano* (MCGH) el cual permita representar los elementos que forman parte del genoma, para lograr de esta manera un mayor entendimiento del dominio.

Es importante resaltar que este objetivo es un gran desafío, debido a que el *genoma humano* conforma un entorno muy *amplio* y *complejo*. Generar un MCGH requiere un marco de trabajo abierto que permita incorporar las nuevas necesidades que vayan surgiendo en el dominio, tomando en consideración los distintos avances médicos que permitan enriquecer o extender dicho modelo.

A través del desarrollo de este trabajo de investigación se pretende demostrar que con el uso de técnicas de modelado conceptual en el dominio genómico se pueden desarrollar *Sistemas de Información Genómicos* eficientes, los cuales permitirían la correcta gestión de toda la información existente.

En el marco de esta Tesis Doctoral se propondrá a partir del MCGH definido varios puntos: a) *argumentar/discutir la necesidad de evolución del modelo*; b) *un caso práctico de extensión del modelo*; y c) *el desarrollo de un prototipo para la gestión de los datos genómicos*. Todo esto con el fin de generar diagnósticos genómicos con datos "*curados*" (adecuadamente revisados) provenientes de distintos repositorios genómicos.

[1] Un *modelo conceptual* es una representación (no sólo) gráfica de una *ontología fundacional* [195] (o de referencia) [196].

1.1 Descripción del Problema

En la actualidad en el entorno bioinformático se pueden encontrar plataformas y herramientas que intentan abordar la gestión de los datos genómicos obtenidos con el paso de los años, pero en la mayoría de los casos son soluciones que permanecen en un estado "*limitado*" e "*inconsistente*" debido a la gran dispersión y heterogeneidad de los datos.

Entender y comprender el comportamiento del genoma humano es una tarea muy difícil, esto principalmente por la amplia cantidad de elementos que participan en el desenvolvimiento de la vida. Es por ello por lo que especialistas en las ramas de biología, genética, y áreas afines (*biomedicina*, *biotecnología*, etc.) unen sus fuerzas para lograr el paso de la medicina tradicional a una medicina completamente personalizada.

Lo importante de todo esto es que los mismos especialistas han reconocido la necesidad de actualizar sus métodos de trabajo y aprovechar todas las ventajas que brindan la aplicación de técnicas de *Sistemas de Información* e *Ingeniería de Software*. Este claro cambio de enfoque ha revolucionado la forma de trabajo y generado la inclusión de expertos en el área de tecnologías de la información con la creación de una amplia gama de grupos y centros de investigación enfocados en el estudio del genoma y sus características, como, por ejemplo, el Proyecto OMIM ("*Online Mendelian Inheritance in Man*"), el Proyecto HapMap, entre otros.

Para entender y visualizar desde una perspectiva holística (*global*) las principales problemáticas del dominio se realizó un estudio exhaustivo, el cual proyectó los siguientes aspectos:

a) La falta de formalización del concepto de "*genoma*"
b) La gran variedad de repositorios genómicos (lo que da lugar a grandes cantidades de datos con estructuras dispersas, heterogéneas y redundantes)
c) La complejidad en la gestión de los datos genómicos

En esta Tesis Doctoral se busca hacer frente a dichos problemas mediante la creación de un *Modelo Conceptual del Genoma Humano*, para demostrar que el uso de este enfoque es esencial para generar *Sistemas de Información Genómicos* que permitan potenciar el

conocimiento existente en el dominio, con el fin de facilitar la tan mencionada "*Medicina de Precisión*".

1.2 Objetivos de la tesis

El objetivo general para resolver en esta Tesis Doctoral es la de proponer y facilitar un Modelo Conceptual del Genoma Humano (MCGH) que permita representar los elementos que participan en el funcionamiento del genoma humano, para lograr de esta forma un mayor entendimiento del dominio.

Para delimitar el problema a resolver, esta Tesis Doctoral se enfoca en cuatro subobjetivos fundamentales:

1) Construir un Modelo Conceptual del Genoma Humano (MCGH) que defina los elementos que componen el *-genoma humano-*.
2) Analizar y evaluar la necesidad de evolución del MCGH
3) Extender el MCGH mediante la integración de los "*Haplotipos*"
4) Implementar un prototipo (*VarSearch*) basado en el MCGH para la gestión de datos genómicos

Estos objetivos conducen a la formulación de las siguientes preguntas objeto de investigación:

> **P1.** *¿En qué ayudaría la definición de un Modelo Conceptual holístico del Genoma Humano para entender el comportamiento y características del genoma humano?*
>
> **P2.** *¿Se podría establecer un Modelo Conceptual del Genoma Humano estático, o debe plantearse una solución abierta (dinámica)?*
>
> **P3.** *¿Es posible extender el Modelo Conceptual del Genoma Humano conservando su definición inicial?*
>
> **P4.** *¿Cuáles son las ventajas de desarrollar una herramienta basada en modelos conceptuales para la gestión de los datos genómicos?*

Las soluciones planteadas para estas preguntas de investigación se especifican a continuación.

1.3 Solución propuesta

Las cuatro contribuciones principales de esta Tesis Doctoral responden, respectivamente, a cada una de las cuatro preguntas de investigación antes planteadas:

1. Generar o diseñar una representación gráfica holística de todos los elementos y conceptos que participan en el funcionamiento del genoma humano mediante la aplicación de técnicas de *modelado conceptual*. Con el uso de este modelo se logrará plasmar el conocimiento existente, y brindar una estructura clara de las iteraciones y conexiones entre los distintos componentes que conforman el genoma (**Capítulo 4**).

2. La solución planteada debe consistir en un modelo conceptual *abierto (dinámico)*, el cual sea extensible para incorporar los nuevos conocimientos que vayan surgiendo en el dominio genómico. Gracias a las tecnologías de NGS cada día se producen grandes cantidades de datos, los cuales generan nuevo conocimiento que deben ser discutidos/analizados para su posterior inclusión en el modelo (**Capítulo 4**).

3. Una vez definido un modelo conceptual holístico del genoma humano, es más fácil visualizar el comportamiento y los elementos participantes en el genoma. Para llevar a cabo la extensión del modelo se tendría que hacer un estudio y análisis del nuevo concepto (caso práctico, "*los haplotipos*") para conocer su impacto y relevancia dentro del modelo. Además, de poder delimitar cual segmento o sección del mismo afectaría y cuál sería su aporte al diagnóstico genómico *-final-* (**Capítulo 5**).

4. El dominio genómico se compone de grandes cantidades de datos dispersos y heterogéneos, que no siempre cuentan con una estructura clara y estándar. Esto provoca una alta susceptibilidad en temas de pérdida de información (conocimiento) por causa de la constante evolución del dominio. Es por ello, que es esencial la creación de herramientas software apoyadas en modelos conceptuales, ya que de esta forma la aplicación consistirá en una proyección del conocimiento ya existente, siempre con la posibilidad de extenderlo con el que tenga que ser generado en versiones

futuras como consecuencia de la aparición de nuevo conocimiento (**Capítulo 6**).

Es importante destacar que para la consecución de los objetivos de esta Tesis Doctoral hemos realizado un estudio del dominio genómico (**Capítulo 2**) y un estado del estado del arte sobre la aplicación de modelado conceptual en el dominio genómico como también el estudio de distintas bases de datos genómicas (**Capítulo 3**).

1.4 Metodología de la Investigación

En esta sección se detallan los distintos aspectos metodológicos en los cuales se fundamenta la presente tesis doctoral. En la subsección 1.4.1 se describe la metodología de investigación utilizada, y en la subsección 1.4.2 su aplicación a esta Tesis Doctoral.

1.4.1 Framework Metodológico

El "*Design Science*[2]" (Ciencia del Diseño) es el diseño y la investigación de artefactos en un contexto. Los artefactos que se estudian están diseñados para *interactuar* con un problema del contexto con el fin de mejorar algo en dicho contexto [2]. En esta tesis doctoral, el artefacto es:

"*Un Modelo Conceptual del Genoma Humano (MCGH) para atender a la heterogeneidad y dispersión de los datos*"

Y el contexto es el: "*Genómico*"

Principalmente, en los proyectos basados en "*Design Science*" se deben considerar dos actividades: *diseño* e *investigación* (ver Figura 1). Estas actividades permiten definir los problemas de diseño (PD) relacionados con el diseño del MCGH, y las preguntas de conocimiento (PC) relacionadas con la búsqueda de conocimiento sobre la interacción entre el MCGH y el contexto en donde se aplica. En la tarea de diseño se clarifica el contexto social de la tesis. En esta se describen las partes

[2] *Design Science:* consiste en estudiar la –*interacción*- de un artefacto en un contexto dado [3].

interesadas (*stakeholders*) que pueden afectar al proyecto o pueden ser afectadas por el mismo:

- El contexto social de esta tesis se compone de distintos tipos de interesados: los usuarios potenciales del MCGH y sus aplicaciones, como, por ejemplo, genetistas, bioinformáticos y laboratorios clínicos que gestionan información genómica, entre otros; los investigadores de esta tesis doctoral: José F. Reyes R. (investigador principal) y Óscar Pastor López (director); el Centro PROS y MESCyT (los cuales proveen el presupuesto para patrocinar la investigación de esta tesis doctoral; las normas sobre salud y protección que podrían beneficiarse de nuevas alternativas de gestión de los datos genómicos.

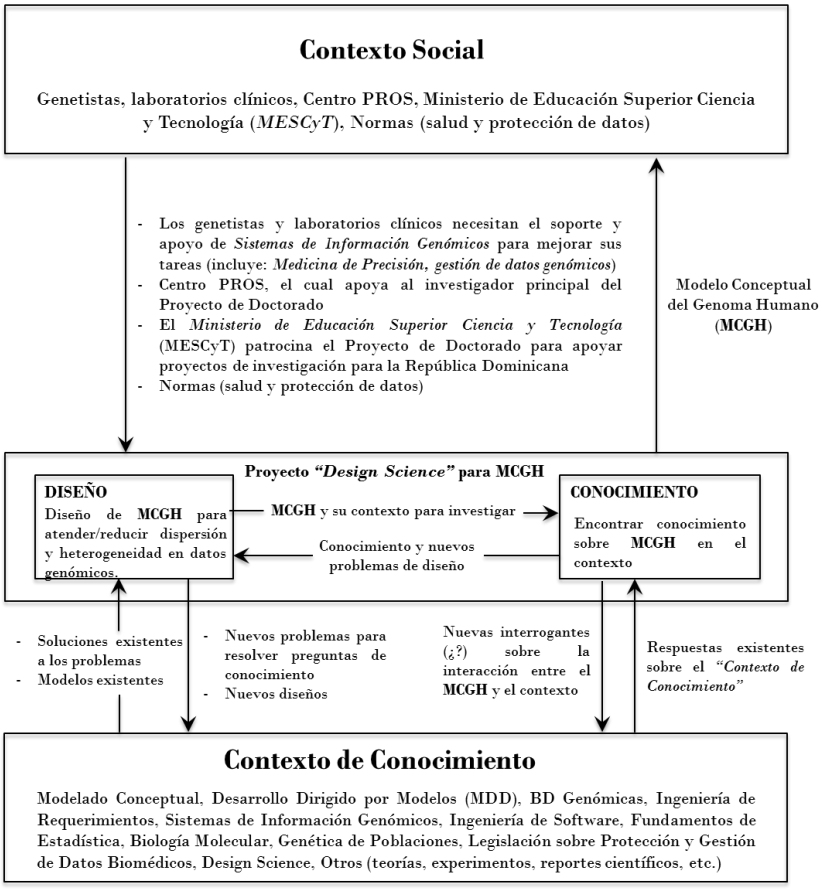


Figura 1. Framework para el "*Design Science*" aplicado al MCGH

En la tarea de investigación se describe el contexto de conocimiento de esta tesis:

- El contexto de conocimiento del MCGH está fundamentado en las técnicas de modelado conceptual y el desarrollo dirigido por modelos. Para el desarrollo de este trabajo se aplicaron tareas de ingeniería de requerimientos e ingeniería de software, análisis de bases de datos genómicas con el fin de construir sistemas de información genómicos. Para el aspecto biológico, fue fundamental el asesoramiento de expertos de las áreas de: biología molecular, genética de poblaciones, estadística aplicada a la biología, legislación sobre tratamiento de datos genómicos. Y para el planteamiento de las actividades de este trabajo se aplicó la metodología de "*Design Science*".

La Figura 1 presenta el framework para el "*Design Science*" aplicado al MCGH. Este framework especifica las relaciones entre la tesis y el contexto tanto social como de conocimiento. El framework junto a los problemas de diseño y preguntas de conocimiento componen las metas/objetivos de la investigación y las preguntas de investigación de esta tesis.

1.4.2 Metodología aplicada a la tesis

El objetivo de esta Tesis Doctoral es resolver el problema: "*Reducir/atender la dispersión de datos heterogéneos en el entorno genómico con el desarrollo de un MCGH para el estudio y explotación de datos más eficientes que den soporte a la medicina personalizada*".

El framework metodológico seleccionado para guiar esta tesis es el propuesto por Wieringa, titulado "*Design Science methodology as set of nested regulative cycles*" [3]. Wieringa propone "*descomponer*" los problemas que se encuentran los investigadores en problemas de *ingeniería* e *investigación*: por lo que un problema principal es considerado como un conjunto de problemas anidados.

La aproximación facilitada para resolver ambos tipos de problemas consiste en seguir un ciclo regulativo (*-regulador-*) basado en cinco tareas: a) *investigación del problema*; b) *diseño*; c) *validación*; d) *implementación*; y e) *evaluación*.

Sin embargo, como se muestra en la Figura 2, dependiendo del tipo de problema, cada problema se aborda con un ciclo regulativo ligeramente distinto. Por ejemplo, un *ciclo de ingeniería* (EC) tiene las tareas de: investigación del problema, especificación de soluciones, validación e implementación de especificaciones y evaluación de la implementación. Mientras que un *ciclo de investigación* (RC) posee las tareas de: la investigación del problema de investigación, el diseño de la investigación, la validación del diseño, la ejecución de la investigación y el análisis de los resultados [4]–[6].

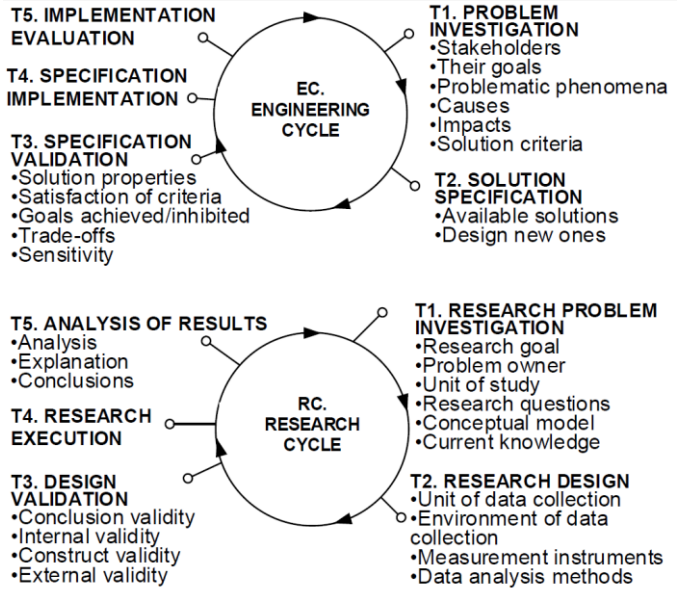


Figura 2. *Design Science* como un ciclo regulativo [3]

Siguiendo la metodología, se planteó el problema como un *problema de ingeniería*, y se desarrollaron las tareas regulativas del *ciclo de ingeniería* (EC).

En la aplicación de esta metodología (Figura 3), se inicia investigando sobre el problema (T1.1) mediante la definición de la motivación de generar un Modelo Conceptual del Genoma Humano (MCGH). Para ello, se debe delimitar los beneficios de aplicar técnicas de modelado conceptual sobre dicho dominio (T1.1.1). A partir de la información obtenida se puede establecer el planteamiento del problema (T1.2). Las tareas siguientes consisten en el estudio y análisis del dominio genómico (T1.2.1) y la aplicación de la metodología de la investigación

(T1.2.2). El siguiente paso consiste en especificar la solución a desarrollar, y para esto fue necesario analizar y/o revisar si alguna propuesta previa del estado del arte satisfacía esta necesidad (T2.1). El objetivo de esta revisión es comprender los problemas que actualmente se encuentran abiertos (T2.1.1) y saber si es necesario proponer una nueva solución. Y como este es el caso, se propone un MCGH y el diseño de un Sistema de Información Genómico (T2.2). Este modelo se analiza y evalúa con el objetivo de discutir las distintas versiones hasta llegar a la versión estable (T2.3) y sus posibles elementos de extensión (crecimiento) (T2.3.1). Tras concluir dichas tareas se procede a integrar toda la información relevante en una única base de datos genómica (T2.4). El siguiente paso consiste en la especificación de la validación. En este paso se debe explicar cómo aplicar la metodología SILE para la correcta obtención de los datos desde los distintos repositorios genómicos (T3.1), y posteriormente validar el MCGH definido mediante el desarrollo de un prototipo (*software*) que permita gestionar los datos genómicos (T3.2). El penúltimo paso consiste en la especificación de la implementación, y en esta se definen los requisitos y resultados esperados, así como las contribuciones brindadas para la comunidad científica (T4.1 y T4.1.1 respectivamente). Por último, la tarea de transferir el resultado de la investigación (T5.1) a la industria queda como un objetivo fuera del alcance de esta Tesis Doctoral.

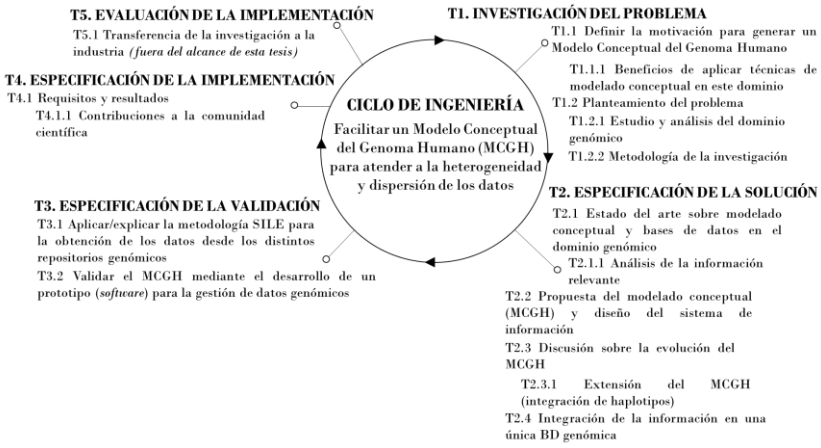


Figura 3. Ciclos regulativos de esta tesis doctoral

1.5 Estructura de la Tesis Doctoral

Una vez planteados los objetivos principales, la estructura del resto de la tesis doctoral se rige en orden a las tareas presentadas en los ciclos regulativos de la metodología. La estructura de la tesis es la siguiente:

- El Capítulo 2 presenta una descripción del dominio genómico, en donde se definen los elementos del genoma humano y su evolución con el paso de los años (1866-2012). También se explican las técnicas de secuenciación basadas en genomas y en exomas. Finalmente, se comentan algunos aspectos relacionados con las tecnologías de secuenciación de última generación (NGS, *Next-Generation Sequencing*) y la Medicina de Precisión (o también conocida "*Medicina Personalizada*").

- El Capítulo 3 muestra el análisis realizado sobre el estado del arte asociado con la presente tesis. Este capítulo se ha dividido de acuerdo con dos líneas de investigación: a) trabajos realizados en el ámbito del modelado conceptual, realizando un mayor énfasis sobre los aplicados en el dominio genómico; y b) trabajos en el entorno de bases de datos genómicas para estudiar los distintos tipos de bases de datos existentes, estructuras y gestión de la información.

- El Capítulo 4 presenta la evolución del Modelo Conceptual del Genoma Humano (MCGH). En este capítulo, se describen las distintas versiones del modelo, y las diferentes decisiones que se realizaron para pasar desde las versiones 1 y 1.1 hasta la versión 2, con el objetivo de cubrir las nuevas necesidades detectadas por causa de la evolución del dominio (mayor entendimiento del dominio).

- El Capítulo 5 presenta la estrategia de integración de los haplotipos en el MCGH. En este capítulo, se justifica, en primer lugar, la necesidad de extender el MCGH. A continuación, se explican los antecedentes sobre los haplotipos detectados en el caso práctico sobre –*la sensibilidad al alcohol*-, así como los trabajos relacionados sobre la representación conceptual de haplotipos en distintos repositorios de datos. De esta forma se realizó la extensión del MCGH con la integración de los haplotipos. Este nuevo modelo fue validado mediante el "*mapping*" de los datos existentes en un conjunto de

repositorios enfocados en: frecuencias alélicas y genotípicas, poblaciones y fenotipos. Por último, se muestra el desarrollo de una base de datos para haplotipos y una comparativa de la evolución de la base de datos según la versión del modelo conceptual.

- El Capítulo 6 introduce la metodología utilizada para la búsqueda de variaciones genéticas en los distintos repositorios genómicos y la Base de Datos del Genoma Humano (HGDB). Además, se presenta el prototipo propuesto ("*VarSearch*") como método de implementación del MCGH previamente definido. En este capítulo también se abordan algunos trabajos relacionados con la búsqueda y anotación de variaciones. Para finalizar se introduce el proyecto *GenesLove.Me* y su aplicación en la Medicina Personalizada.

- El Capítulo 7 presenta las conclusiones y contribuciones principales de la presente Tesis Doctoral, las líneas de trabajo futuro y el impacto de la tesis (por ejemplo, *publicaciones* y *proyectos académicos realizados*).

CAPÍTULO 2

Dominio Genómico

Hoy en día, el entendimiento del *genoma humano* es un gran desafío, probablemente el desafío más importante al que se enfrenta la humanidad en este siglo XXI. Y esto se debe en gran parte a la complejidad que representa este dominio. En este capítulo se presenta el contexto de aplicación de esta tesis, la cual tiene como objetivo plantear las nociones y fundamentos esenciales del dominio. Es un entorno de aplicación especialmente interesante en dos planos fundamentales: en primer lugar, por las implicaciones sociológicas que implica plantearse la posibilidad de entender el lenguaje de la vida. Y, en segundo lugar, desde una perspectiva más práctica de aplicación en el ámbito clínico, debido a su repercusión en la generación de diagnósticos genómicos, los cuales juegan un papel importante dentro de la *Medicina de Precisión*.

Todo ello dentro de un marco de trabajo centrado en el uso del *Modelado Conceptual* como estrategia esencial de búsqueda de soluciones. Pero antes de hablar de soluciones, entendamos bien el problema. A ello se va a dedicar este capítulo.

Con ese objetivo en mente, en primer lugar, en la Sección 2.1, se introducen los conceptos relacionados con el genoma humano y una cronología de la evolución del dominio. La Sección 2.2 explica en qué consiste la secuenciación de genomas y se introduce brevemente una descripción de las pruebas genéticas (sus usos y tipos), la secuenciación de exomas y las distintas tecnologías de secuenciación (por ejemplo, "NGS"). En la Sección 2.3 se introduce el concepto de medicina de precisión y sus beneficios para la salud. Por último, en la Sección 2.4 se presentan las conclusiones del capítulo. Resultados asociados a este capítulo se encuentran publicados en los siguientes trabajos [7] y [8].

2.1 Genoma Humano

¿Qué es la Genómica? ¿Qué es el genoma?

Rodríguez M. define la *genómica* como la sub-disciplina de la genética interesada en la descripción y análisis molecular de genomas completos. Esta se suele subdividir en dos grandes áreas:

> 1) la *genómica estructural*, que se ocupa de la caracterización de la naturaleza física de los genomas, y
> 2) la *genómica funcional*, cuyo objetivo último es ubicar todos los elementos integrantes de un genoma dentro de una estructura funcional, tanto en el sentido más tradicional de determinar la función de cada uno de los elementos componentes de un genoma (las proteínas codificadas, los elementos reguladores, estructurales, etc.) como en el sentido más general de determinar el papel que cada uno de estos elementos desempeña en el funcionamiento global del organismo [9].

El *Instituto Nacional de Investigación del Genoma Humano* (NHGRI) define el genoma como:

"*Una colección completa de ácido desoxirribonucleico (ADN) de un organismo, o sea un compuesto químico que contiene las instrucciones genéticas necesarias para desarrollar y dirigir las actividades de todo organismo. Las moléculas del ADN están conformadas por dos hélices torcidas y emparejadas. Cada hélice está formada por cuatro unidades químicas, denominadas bases nucleótidas. Las bases son adenina (A), timina (T), guanina (G) y citosina (C). Las bases en las hélices opuestas se emparejan específicamente; una A siempre se empareja con una T, y una C siempre con una G.*

El genoma humano contiene aproximadamente 3.000 millones de estos pares de bases, los cuales se encuentran en los 23 pares de cromosomas dentro del núcleo de todas nuestras células. Cada cromosoma contiene cientos de miles de genes, los cuales tienen las instrucciones para sintetizar proteínas. Cada uno de los 30.000 genes estimados en el genoma humano produce un promedio de tres proteínas." [10].

El genoma humano es un tema de gran interés, por el carácter disruptivo de los resultados que se pueden obtener a través de la manipulación del conocimiento genómico generado. Ese conocimiento va cada día en aumento y facilita una amplia gama de alternativas que pueden mejorar de una manera desconocida hasta ahora la calidad de vida de los individuos. De acuerdo con la definición del NHGRI, se puede deducir que para lograr un mayor entendimiento del *genoma humano* se requiere:

1. Estudiar y ampliar el conocimiento genómico existente
2. Gestionar grandes cantidades de información

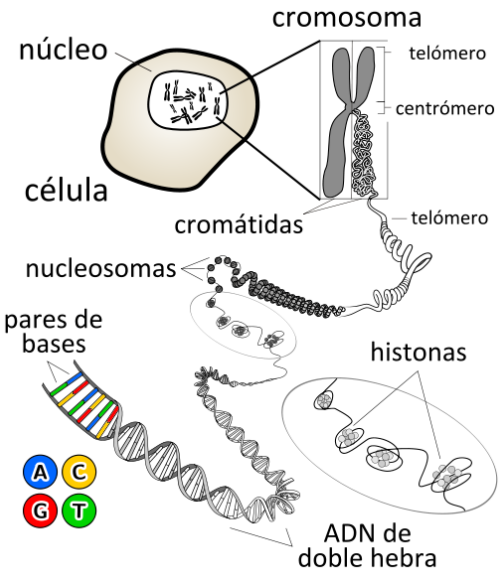


Figura 4. Elementos del cuerpo humano [11]:
El cuerpo humano está formado por trillones de células, las cuales tienen en su núcleo 23 pares de cromosomas (la mitad son del padre y la otra mitad de la madre). Los cromosomas contienen aproximadamente 3 millones de pares de bases (nucleótidos) en el ADN. El ADN y las histonas (son las proteínas sobre las que se enrolla el ADN) se enrollan formando de esta manera los cromosomas. Así, cada proceso biológico tiene un objetivo fundamental en el desenvolvimiento diario en la vida del individuo.

Para ello es importante aplicar métodos/técnicas que ayuden y faciliten el manejo de una gran variedad de conceptos *relevantes* (Figura 4), los cuales tienen una participación vital en uno o varios procesos biológicos.

El volumen de datos asociado a la gestión de datos genómicos es enorme, y hace necesaria la utilización de mecanismos potentes que permitan explotar los datos de manera eficiente. En un dominio de trabajo claramente asociado a la problemática "*Big Data*", se debe necesariamente considerar el uso de tecnologías *Big Data* avanzadas. *¿Por qué?* Cuando se estudian las características del Big Data (4V), estas se encuentran involucradas en este dominio, tal cual como se presenta en la Tabla 1 [12]:

- *Volumen* → datos a escala, factor número uno del dominio: las grandes cantidades de información genómica.
- *Variedad* → datos de muchas formas, múltiples estructuras de datos (forma de representación).
- *Velocidad* → datos "*in motion*", datos que se mantienen en movimiento, en constante evolución.
- *Veracidad* → datos "*uncertainty*", datos facilitados que representan altos niveles de incertidumbre debido a los problemas de heterogeneidad, dispersión, redundancia, entre otros en los distintos repositorios de datos genómicos.

Tabla 1. Cuatro dominios de Big Data en 2025. En cada uno de los cuatro dominios se presentan las necesidades anuales de almacenamiento y computación proyectadas a lo largo del ciclo de vida de los datos. Fuente: *Big Data: Astronomical or Genomical?* [12]

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

A título de ejemplo, y solo a nivel de volumen, a nivel de almacenamiento de las muestras de un individuo –*genoma de una persona*-, se puede decir que el espacio utilizado (promedio) para guardarlo está entre 3,5GB± a 5GB±.

¡Puede fácilmente imaginarse la complejidad asociada a la gestión de un banco de datos genómico que incorporara por ejemplo los genomas de todos los ciudadanos de la *Comunidad Valenciana*, o del *Estado Español*, o de la *Unión Europea*… y podríamos continuar ampliando el contexto de trabajo!

Con el fin de buscar e implementar soluciones sobre este contexto surgió el *Proyecto del Genoma Humano*, el cual tenía como objetivo principal determinar la secuencia completa del genoma humano, para alcanzar una serie de objetivos técnicos y éticos [13]:

a) desarrollo de técnicas de secuenciación rápidas y automatizadas;

b) desarrollo de bases de datos y programas informáticos para manejar y analizar el enorme volumen de información generado;

c) discusión de aspectos éticos y legales relacionados con el genoma humano; y

d) desarrollo de políticas adecuadas para garantizar: la privacidad de los datos genéticos, el respeto a la diversidad genética y la correcta utilización del diagnóstico genético.

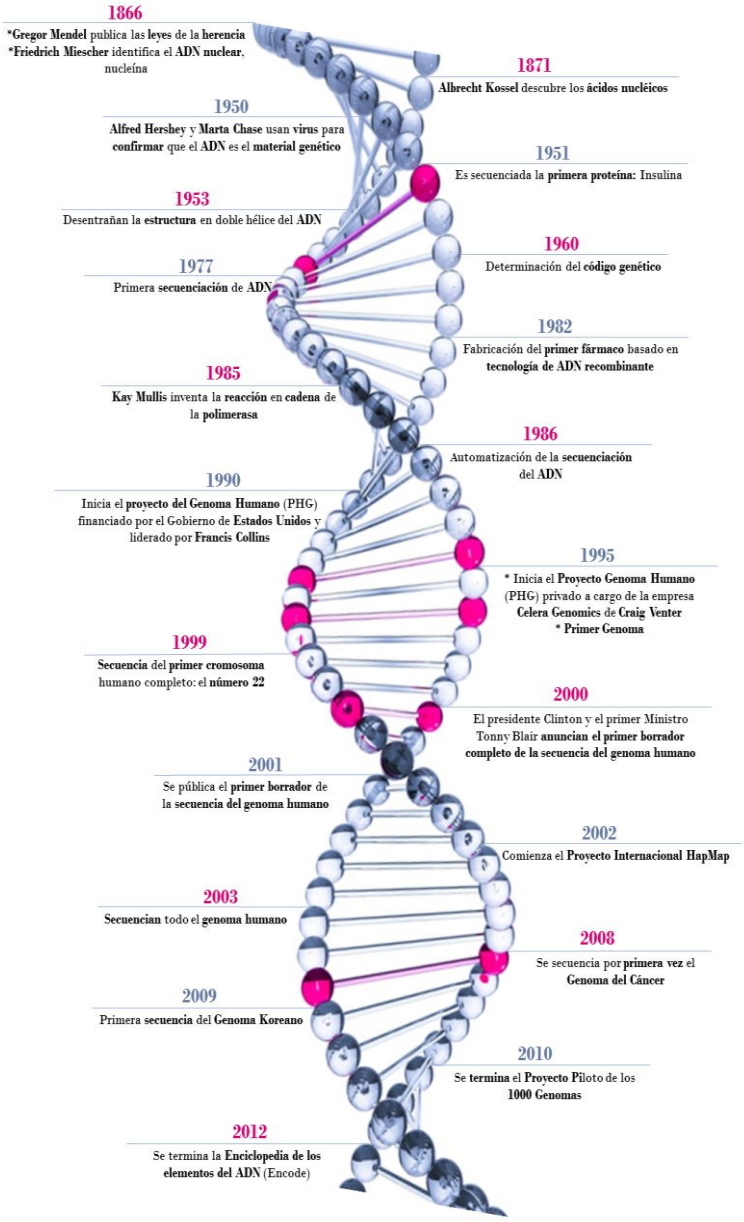


Figura 5. Cronología del genoma humano (1866-2012)

Para entender la secuencia histórica de avances asociados al conocimiento del genoma, la Figura 5 muestra una *cronología del*

genoma humano realizada a partir de diversos estudios [14] [15], en donde se puede visualizar la evolución del dominio desde sus inicios en el año 1866, cuando Gregor Mendel publicó las leyes de la herencia hasta el año 2012, con la finalización de la Enciclopedia de los elementos de ADN (*Encode*). En la figura siguiente (Figura 6) se puede ver la noticia reportada por el diario "*El País*" en junio del año 2000 sobre el anuncio del primer borrador completo de la secuencia del genoma humano [16].


ESPAÑA | AMÉRICA | BRASIL | CATALUÑA NEWSLETTER SUSCRÍBETE

≡ EL PAÍS SOCIEDAD

LA REVOLUCIÓN GENÉTICA

El descifrado del genoma humano abre una nueva era en la lucha contra las enfermedades

Celera Genomics y el proyecto público descifran el 'libro de la vida' y acuerdan trabajar juntos - Clinton y Blair piden un uso responsable de la información - El 99,8% de los datos genéticos son comunes a todas las personas

EL PAÍS

Washington - 27 JUN 2000

El 26 de junio de 2000 es ya una fecha para la historia de la humanidad. Tras 10 años de intensa investigación, el genoma humano, considerado el auténtico libro de la vida, se dio ayer por descifrado en sus partes esenciales. El revolucionario logro, que abre una nueva era en la lucha contra las enfermedades, fue anunciado consecutivamente en China, Japón, Francia, Alemania, Londres y Estados Unidos. Para conseguir este hito, que corona un siglo de investigación biológica, el proyecto público internacional y el privado de la empresa estadounidense PE Celera Genomics abandonaron la pugna que mantenían y decidieron anunciarlo conjuntamente en la Casa Blanca, en una ceremonia presidida por el presidente Bill Clinton.

Figura 6. Noticia diario "*El País*" (27-junio-2000)

En la actualidad, se continúa uniendo esfuerzos con el fin de seguir avanzando en este contexto debido especialmente a su repercusión directa en un entorno tan sensible como el clínico, ligado a la nueva *medicina personalizada* o de *precisión*. El estudio de los *genes* ha demostrado que estos juegan un papel importante en las distintas enfermedades. Y muchos de los factores que se analizan en el dominio genómico indican que las principales causas de mortalidad poseen un origen genético, como, por ejemplo, la *diabetes*, *cáncer*, *enfermedades cardiacas*, entre muchas otras [13].

Gracias a los proyectos de investigación en este campo, y a la colaboración de una amplia gama de especialistas/expertos (por ejemplo, *biólogos*, *genetistas*, *tecnólogos*, etc.) se van alcanzando muchos

de los objetivos planteados en este dominio, además de despejar un camino tan desconocido e incierto mediante el trabajo colaborativo y compartido. Todavía faltan mecanismos y soluciones más viables, pero lo importante de todo el recorrido realizado es poder trazar pautas que permitan llegar al objetivo de llegar a entender el genoma humano de forma sistematizada.

2.2 Secuenciación de Genomas

Desde 1977, cuando se inició la secuenciación del ADN y surgió el desarrollo de *software* para el análisis de los datos de forma rápida, se inició una carrera que culminó a los pocos años con la publicación de la primera secuencia genética -*completa*- de un organismo [17].

La secuenciación de genomas tiene innumerables beneficios, como las pruebas genéticas (*mencionadas en el punto anterior*), ya que permiten identificar mutaciones o alteraciones en los genes, lo cual posee un gran valor para la medicina clínica, por su impacto en el diagnóstico precoz de las enfermedades [18].

La *secuenciación del genoma* consiste en determinar el orden exacto de los pares de bases en un segmento de ADN. Los cromosomas humanos tienen entre 50.000.000 a 300.000.000 pares de bases. Debido a que las bases existen en pares, y la identidad de una de las bases en el par determina el otro miembro del par, los científicos no tienen que presentar las dos bases del par. El principal método utilizado por el PGH (*Proyecto Genoma Humano*) para producir la versión final del código genético humano se basa en un mapa, o en una secuencia basada en BAC, que es el acrónimo en inglés de "*cromosoma artificial bacteriano*". El ADN humano es fragmentado en piezas relativamente grandes, pero de un tamaño manejable (entre 150.000 y 200.000 pares de bases). Los fragmentos son clonados en bacterias, las cuales almacenan y replican el ADN humano para que así pueda ser preparado en cantidades lo suficientemente grandes como para secuenciarlo. Si se los escoge cuidadosamente para minimizar las superposiciones, se necesita unos 20.000 clones BAC diferentes para abarcar los 3.000 millones de pares de bases del genoma humano. A la colección de clones BAC que contienen todo el genoma humano se la denomina una *"biblioteca BAC*" [10] [19].

La bioinformática mediante sus múltiples ramas (por ejemplo, *investigaciones sobre análisis de secuencias*, *anotación del genoma*, *análisis de mutaciones en cáncer*, *sistemas biológicos comparativos*, *acoplamiento proteína-proteína*, etc.) da la posibilidad de aplicar el conocimiento obtenido en todos estos años a *la gestión de bases de datos biológicas*, *procesos metabólicos*, *genética de poblaciones*, *inteligencia artificial*, entre otros [17], [20].

2.2.1 Pruebas Genéticas

Un caso práctico de los avances y contribuciones alcanzados por el contexto bioinformático, son las *pruebas genéticas*[3].

Las pruebas genéticas han sido esenciales para ayudar a prevenir y tratar enfermedades de origen genético [21]. Para ello, se emplean métodos de laboratorio con el fin de estudiar los genes (que incluyen las instrucciones del ADN heredado de la madre y del padre). Estas pruebas se utilizan para identificar mayores riesgos de problemas de salud, elegir tratamientos o evaluar respuestas a tratamientos [22].

El diagnóstico de enfermedades genéticas implica un examen clínico integral, el cual consta de tres elementos principales:
1) *examen físico*;
2) *antecedentes familiares detallados*; y
3) *pruebas clínicas y de laboratorio* [21]

En el capítulo 2 sobre *Diagnóstico de una enfermedad genética* del libro "*Cómo entender la genética: Una guía para pacientes y profesionales médicos en la región de Nueva York y el Atlántico Medio*" [21], y en otros trabajos asociados [23]–[25] se pueden encontrar los distintos usos y tipos de las pruebas genéticas con sus respectivas definiciones, las cuales se presentan brevemente a continuación:

a) **Usos de las pruebas genéticas:**
   - *La detección sistemática o tamizaje neonatal* es la prueba genética más realizada. En muchos países se ha adoptado como una práctica sistemática, cuyo objetivo es la

[3] Las pruebas genéticas son exámenes de sangre y otros tejidos para detectar trastornos genéticos.

detección temprana y el tratamiento de los recién nacidos -*enfermos*-, aunque todavía presintomáticos [26].

- *Las pruebas de detección de portadores* pueden ayudar a las parejas a saber si son portadores, y en tal caso, el riesgo de transmisión a sus hijos, del alelo (*variante de un mismo gen*) de una enfermedad recesiva como la fibrosis quística, la anemia falciforme o la enfermedad de Tay-Sachs.
- *Las pruebas de diagnóstico prenatal* sirven para detectar modificaciones en los genes o los cromosomas de un feto. Este tipo de pruebas se recomienda a las parejas que presentan un mayor riesgo de tener un bebé con un trastorno genético o cromosómico identificable.
- Las pruebas genéticas pueden usarse para confirmar un *diagnóstico* en un individuo que presenta ciertos síntomas o para monitorear el pronóstico de una enfermedad o la respuesta a un tratamiento médico.
- *Las pruebas predictivas o de predisposición* sirven para identificar a las personas con riesgo de una enfermedad antes de la aparición de los síntomas. Estas pruebas son muy útiles cuando una persona tiene antecedentes familiares de una enfermedad en particular y existe un método de intervención disponible para prevenir la aparición de dicha enfermedad o para minimizar su gravedad. Las pruebas predictivas sirven para identificar las mutaciones que aumentan el riesgo que una persona desarrolle una enfermedad de origen genético, como es el caso de algunos tipos de cáncer.

b) Tipos de pruebas genéticas: el tipo de prueba depende del tipo de anomalía que se esté evaluando.

- *Pruebas citogenéticas*: La citogenética implica la evaluación de todos los cromosomas para detectar anomalías. Los cromosomas de las células humanas en división pueden analizarse sin problema bajo un microscopio. Los glóbulos blancos, particularmente los linfocitos T, son las células más disponibles y más accesibles para análisis citogenéticos ya que pueden obtenerse fácilmente de la sangre y se dividen rápidamente en un cultivo celular. Las células de los tejidos como la médula ósea (para la leucemia), el líquido

amniótico (para el diagnóstico prenatal) y las biopsias de otros tejidos también se pueden cultivar para realizar análisis citogenéticos.

- *Pruebas bioquímicas*: La enorme cantidad de reacciones bioquímicas que ocurre a diario en las células requiere diferentes tipos de proteínas. Por lo tanto, hay diferentes tipos de proteínas como enzimas, transportadores, proteínas estructurales, proteínas reguladoras y hormonas que cumplen diferentes funciones. La mutación de cualquier tipo de proteína puede causar una enfermedad si esta mutación no permite que la proteína funcione correctamente.
- *Pruebas moleculares*: Para las pequeñas mutaciones de ADN, las pruebas directas del ADN suelen ser el método más eficaz, especialmente si la función de la proteína es desconocida y no se puede desarrollar una prueba bioquímica. Las pruebas de ADN se pueden realizar en cualquier muestra de tejido, incluso con muestras muy pequeñas. Las pruebas moleculares representan un gran desafío ya que algunas enfermedades genéticas pueden estar relacionadas con un gran número de mutaciones diferentes.

La bioinformática se dedica a contribuir en el crecimiento y evolución del conocimiento genómico, logrando de esta forma que lo que hace unos años era un mito o algo inalcanzable, hoy en día pueda ser una realidad. Cada vez está más cerca la completa *medicina personalizada*, y paso a paso se está cambiando de la tradicional "*prescripción empírica*", la cual corre el riesgo de repercutir en pérdida de tiempo y dinero, a una "*prescripción personalizada*", la cual saca provecho del conocimiento genómico existente, y abre paso a la *farmacogenómica* [27], la cual proyecta un tratamiento conforme a las necesidades individuales, lo que garantiza una mayor eficacia en el tratamiento.

La gestión de los datos genómicos requiere la aplicación de mecanismos de seguridad que puedan garantizar la *disponibilidad*, *integridad*, *confidencialidad (probablemente la más importante)*, *trazabilidad* y *autenticidad* de la información, así como también tomar en consideración todos los aspectos legales que repercuten sobre estos datos (*sensibles*). Solo en este ámbito cabe destacar la complejidad derivada de la necesidad de legislar un asunto tan novedoso y en

evolución constante. A título de ejemplo, las siguientes normas legales están directamente implicadas en este problema:

- *Ley Orgánica 15/1999*, de 13 de diciembre, de Protección de Datos de Carácter Personal [28].
- *Ley 14/2007*, de 3 de julio, de Investigación biomédica [29], [30].
- *Ley 14/1986*, de 25 de abril, General de Sanidad [31].

2.2.2 ¿Qué es la secuenciación de exomas?

"*La secuenciación completa del exoma (Secuenciación de Exoma Completo -Whole Exome Sequencing, WES por sus siglas en inglés-) se puede utilizar para identificar la base molecular del trastorno en un individuo afectado. La prueba exoma es diferente de otros tipos de pruebas de diagnóstico genético en términos del número de genes que son secuenciados simultáneamente. La prueba exoma se dirige a las regiones codificantes de los genes. Estas regiones específicas de genes de un individuo, llamados exones, se capturan y se secuencian utilizando secuenciación masiva. La secuencia de un individuo se compara a continuación con secuencias de referencia publicadas en bases de datos internacionales lo que ayuda a identificar las variantes causales que podrían explicar el desorden en el paciente afectado* [32]".

Este estudio se realiza en dos etapas. En primer lugar, se realiza la secuenciación del exoma. En la segunda etapa se realiza el análisis de las secuencias en donde todas las variantes de la muestra se identifican y vinculan las variantes con la enfermedad.

La ventaja fundamental de la secuenciación completa del exoma es poder tener la secuencia de todos los genes funcionales a un costo mucho menor que uno a uno.

Otra ventaja de la secuenciación completa del exoma es que todos los exones en todos los genes se secuencian. Así que la prueba es muy amplia y tiene un rendimiento diagnóstico superior en comparación con los métodos convencionales de otras pruebas genéticas de análisis de genes individuales, paneles de mutaciones frecuentes o utilizando arrays CGH (*Comparative Genomic Hybridization*) [32]–[34].

2.2.3 Tecnologías de Secuenciación

Gracias a las técnicas de NGS (*Next-Generation Sequencing*) se ha logrado reducir los costes de secuenciación, permitiendo que la investigación clínica y científica pueda seguir avanzando, y que cada día se vaya mejorando la comprensión del genoma. La evolución de estas tecnologías permite que hoy en día se continúen alcanzando logros científicos importantes. Cada vez es mayor el conocimiento sobre nuevos *genes/variaciones* que impulsan la resolución de las bases genéticas asociadas a enfermedades de origen genético. En el trabajo titulado "*Ten years of next-generation sequencing technology*" de *van Dijk et. al.* (2014) se presenta la Figura 7 (extraída) con una comparativa de la evolución de las distintas plataformas de secuenciación de alto rendimiento [35].

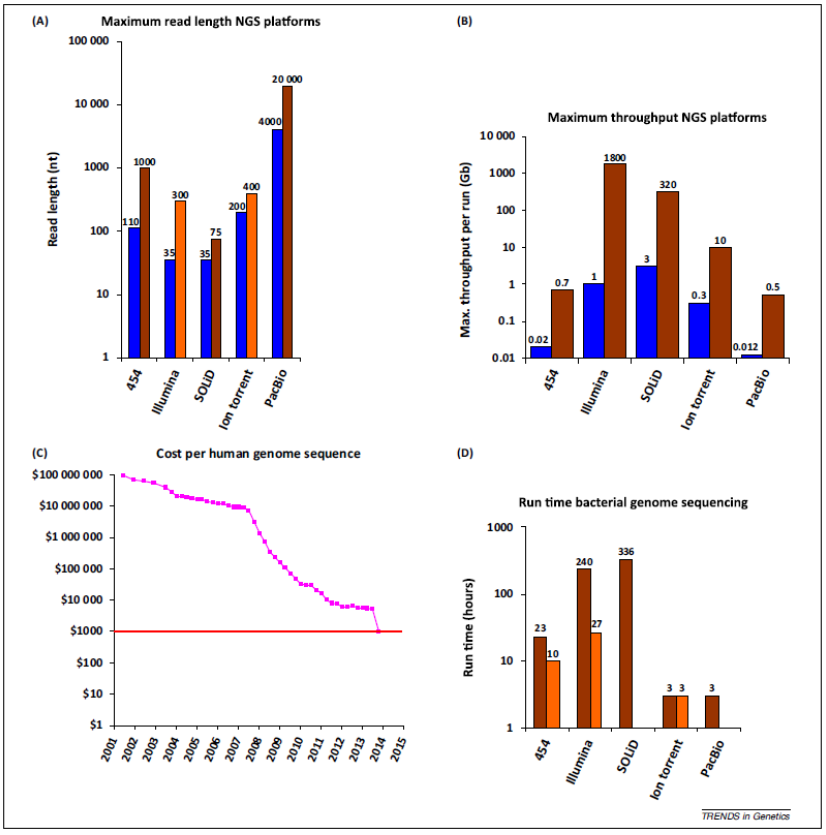


Figura 7. Evolución plataformas de secuenciación de alto rendimiento [35]

La gráfica (A) representa la longitud máxima de lectura de los primeros instrumentos de secuenciación comercialmente disponibles por *454*, *Illumina/Solexa*, *SOLiD*, *Ion Torrent*, y *Pacific Biosciences (PacBio)*. La gráfica (B) representa el rendimiento máximo de los primeros instrumentos de secuenciación comercialmente disponibles (barras azules) contra el rendimiento máximo actual (barras de color naranja oscuro). La gráfica (C) representa la evolución del coste de secuenciación de un genoma humano desde 2001-2014, es importante resaltar que dichos costos han disminuido drásticamente debido a la aparición de las NGS y sus posteriores mejoras. Por último, la gráfica (D) representa los tiempos para completar una ejecución típica para secuenciar un genoma bacteriano utilizando los grandes instrumentos de los distintos fabricantes (barras de color naranja oscuro) contra las máquinas nuevas -*actuales*- [35].

Para finalizar este apartado, se presentan dos tablas comparativas:

1) *comparativa entre las diferentes plataformas de secuenciación* (Tabla 2); y

2) las *ventajas y desventajas de las diferentes estrategias de secuenciación* (Tabla 3) reportadas en el artículo "*Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal*" por B. Rodríguez-Santiago y L. Armengol [36].

Tabla 2. Comparativa entre diferentes plataformas de secuenciación [36]

Tabla 1 – Comparativa entre diferentes plataformas de secuenciación

Plataforma	Tiempo carrera[a]	*Reads*/carrera (en millones)	Bases/*read*[b]	Rendimiento (Mb/carrera)
3730xl (capilares, no NGS)	2 h	0,000096	650	0.06
Ion Torrent (chip 314)	2 h	0,10	100	>10
454 GS Jr. Titanium	10 h	0,10	400	50
Starlight[g]	?	~0,01	> 1.000	?
PacBio RS	0,5 – 2 h	0,01	860–1.100	5-10
454 FLX Titanium	10 h	1	400	500
454 FLX+[c]	18–20 h.	1	700	900
Ion Torrent (chip 316)	2 h	1	> 100	> 100
Helicos[d]	N/A	800	35	28.000
Ion Torrent (chip 318)	2 h	4-8	>100	>1.000
Illumina MiSeq[g]	26 h	3,4	150+150	1.020
Illumina iScanSQ	8 días	250	100+100	50.000
Illumina GAIIx	14 días	320	150+150	96.000
SOLiD – 4	12 días	> 840[e]	50+35	71.400
Illumina HiSeq 1000	8 días	500	100+100	100.000
Illumina HiSeq 2000	8 días	1.000	100+100	200.000
SOLiD – 5500 (PI)[g]	8 días	> 700[e]	75+35	77.000
SOLiD – 5500xl (4hq)[g]	8 días	> 1.410[e]	75+35	155.100
Illumina HiSeq 2000 – v3[f g]	10 días	≤ 3.000	100+100	≤ 600.000

h: horas; Mb: megabases.
~Valor probablemente derivado de información no publicada indisponible en mayo de 2011.
Adaptado de Glenn, 2011[59].
[a] Tiempo necesario en el instrumento para conseguir la longitud máxima de *read*.
[b] Longitud promedio para los *reads* de alta calidad.
[c] Actualización del instrumento FLX, verano 2011.
[d] Instrumentos y reactivos no se pueden comprar ya, solo ofrecen servicios.
[e] *Reads* alineables (número crudo de *reads* de alta calidad).
[f] Reactivos y *software* TruSeq v3 anunciados, *reads* y rendimiento son la mitad que el HiSeq1000.
[g] Información basada solamente en datos de la compañía (datos independentes todavía no disponibles).

Tabla 3. Ventajas y desventajas: distintas estrategias de secuenciación [36]

Tabla 2 – Ventajas y desventajas de diferentes estrategias de secuenciación

Estrategia	Ventajas	Desventajas
Secuenciación Sanger gen a gen	Muy precisa Coste bajo por exón Alto rendimiento	Rendimiento diagnóstico bajo en trastornos genéticos heterogéneos
NGS dirigida a *loci* específicos de enfermedad	Se puede optimizar Riesgo bajo de hallazgos con significado incierto Facilidad de manejo de datos Facilidad en la interpretación	Diseño y rediseño necesarios para nuevos *loci* Experimentos diferentes para cada enfermedad diferente
Secuenciación del exoma	Estudios no sesgados Mismo experimento para cualquier enfermedad Cada experimento contribuye a la interpretación de otros experimentos Puede ser reinterpretado Buena para detección de bajo mosaicismo	Sesgos en la secuenciación No da información de regiones no codificantes Posibilidad de hallazgos con significado incierto
Secuenciación del genoma completo	Sin sesgos en la secuenciación La mejor para detección de variantes estructurales	Manipulación de datos compleja Interpretación compleja Posibilidad de hallazgos con significado incierto

Para cerrar esta sección, se plantea una reflexión que está en el "*genoma*" de esta Tesis Doctoral. Nos enfrentamos a un problema inmensamente complejo, que involucra a un volumen enorme de datos cuyo significado biológico se va descubriendo y enriqueciendo día a día, y que requiere de un soporte tecnológico sofisticado pero que está también en un proceso de mejora continuada.

Cada vez hay más datos, y de esos datos hay que inferir aquellos patrones de información que sean realmente significativos para resolver el problema de fondo que abordamos en este trabajo: el entendimiento del genoma.

El argumento esencial de esta Tesis Doctoral es que ese desafío tan complejo solo puede ser resuelto mediante el uso de técnicas de modelado conceptual, que permitan estructurar el conjunto mínimo de información válida para poder gestionar esa complejidad de datos y para poder elaborar modelos de conocimiento claros y eficientes. Este va a ser el punto central de la solución presentada en esta Tesis Doctoral en los capítulos próximos.

2.3 *Background*: Medicina de Precisión

El contexto social de este trabajo se basa en la "*Medicina de Precisión*". Este enfoque emergente busca el tratamiento y la prevención de las enfermedades con el objetivo de cambiar lo anteriormente denominado como "*Practicas de Medicina Tradicional*".

La medicina de precisión tiene en cuenta la variabilidad en los *genes*, el *medio ambiente* y el *estilo de vida* de cada persona para proporcionar tratamientos individualizados y la prevención de enfermedades [37].

Hablar de medicina de precisión es pues hablar de una forma diferente de entender y tratar al paciente, que permite a los médicos identificar una enfermedad y seleccionar los tratamientos que sean más propensos a ayudar a los pacientes de acuerdo a su diagnóstico genético relacionado con la enfermedad sufrida por él mismo (también se le conoce como "*Medicina Personalizada*") [38].

Figura 8. El uso de cambios genéticos en el tumor de un paciente para determinar su tratamiento se conoce como medicina de precisión [38]

En términos prácticos, *Fowler* (2014) describe que gracias a los avances en la genómica se permitirá proporcionar información adicional sobre las enfermedades, explicando cuáles son las personas con mayor probabilidad de padecer estas enfermedades y cómo aplicar un tratamiento más exitoso para cada individuo [39] (un ejemplo se puede visualizar en la Figura 8). Por ejemplo, aunque existen muchas causas de cáncer de pulmón, sólo las personas que tienen una determinada mutación en el gen "*EGFR*" responden al tratamiento con "*tyrosine kinase inhibitors*" [40]. Incluso cuando se conoce la causa de una

enfermedad, las diferentes variantes genéticas pueden afectar la eficacia del tratamiento alterando la forma en que se metabolizan los fármacos o aumentando la probabilidad de eventos adversos [41].

Otra ventaja de la medicina de precisión es que puede combinar temas de *epidemiología*, *clínica*, *genómica* y *preferencias personales* en una revolución audaz para prevenir y tratar enfermedades.

La investigación en este campo puede ayudar a mejorar las tasas de supervivencia en enfermedades como el cáncer, o podría encontrarse nuevas opciones de tratamiento para enfermedades raras [42] para las cuales no existen un tratamiento específico.

La prevención también tiene un importante impacto no sólo en evitar el desarrollo de enfermedades, sino también en la reducción de los costos médicos. Para lograr estos objetivos –*prevención*, *tratamiento*, *conocimiento*- se deben combinar las distintas tecnologías lanzadas en el entorno genómico e informático.

2.4 Conclusiones

Gracias a los estudios sobre el genoma humano se han logrado obtener avances transcendentales que repercuten en la mejora de tratamientos y prevención de enfermedades de origen genético. Desde que el primer genoma humano secuenciado fue publicado (*completo*) en 2003, ha habido un progreso asombroso en cuanto a velocidad y costo de secuenciación. Esta tarea se tomó 13 años y una inversión aproximada de tres mil millones de dólares.

Con la introducción de las tecnologías NGS la adquisición de los datos ha sido un desafío superado, ahora bien, el almacenamiento, la gestión y la interpretación de los datos se ha convertido en el principal problema dentro de este dominio, y para su resolución se requiere la colaboración de un amplio grupo de especialistas, entre los que se incluyen biólogos computacionales, informáticos, consejeros genéticos o patólogos [43].

La medicina de precisión ha revolucionado la forma en que históricamente se ha entendido la "*medicina*". Este nuevo contexto (práctico) requiere la integración de técnicas de ingeniería de software que permitan capturar y estructurar toda la información relevante del dominio, en donde se habla de una amplia complejidad por ser un contexto cambiante o en constante evolución. Presentado el dominio de trabajo, surge un problema fundamental al que se enfrenta este trabajo de Tesis Doctoral. Los avances en tecnologías de secuenciación genómica hacen que las posibilidades de obtener más y más información de interés aumenten constantemente.

El tratamiento de esa inmensa cantidad de datos hace imprescindible disponer de mecanismos de gestión de datos avanzada que permitan estructurar esos datos para poder interpretarlos y explotarlos de una forma efectiva y eficiente. El objetivo fundamental en esa dirección pasa por plantear una *Bioinformática dirigida por Modelos Conceptuales*. La caracterización de un *Modelo Conceptual del Genoma Humano* como herramienta esencial para ese proceso de gestión efectiva y eficiente de datos genómicos va a ser el objetivo esencial discutido como solución al problema en los capítulos siguientes. Antes de presentar ese Modelo Conceptual, el próximo capítulo tiene como objetivo delimitar qué trabajos se pueden encontrar actualmente en esa dirección, para contextualizar adecuadamente la solución presentada en esta Tesis Doctoral.

CAPÍTULO 3

Estado del Arte

En el marco de la Tesis Doctoral, se presenta un *Modelo Conceptual del Genoma Humano* (MCGH) y su respectiva evolución de acuerdo al conocimiento generado con el paso de los años en el dominio genómico. Además, como ejemplo de su constante y necesaria evolución, se propone la integración de los haplotipos, los cuales repercuten en temas de: *frecuencias*, *poblaciones* y *factores estadísticos*. A partir de este MCGH se presentan las principales contribuciones de este trabajo. Por lo tanto, en esta sección 3.1, se ha realizado un breve resumen sobre los trabajos más relevantes en *modelado conceptual para el dominio genómico*.

El uso de modelado conceptual está ganando impulso como un enfoque de desarrollo de software en el campo médico, con el objetivo de mejorar en gran medida el trabajo realizado por *genetistas*, *científicos de laboratorio* y *médicos*. A continuación, en la sección 3.2, se presenta un breve resumen sobre las bases de datos genómicas, y su repercusión en la gestión de los datos genómicos. Además, en las siguientes subsecciones se presenta una lista de hasta 22 bases de datos genómicas –*consideradas populares*-, y de gran repercusión en el estudio de enfermedades de origen genético. Es importante conocer la amplia

diversidad de bases de datos genómicas que existen actualmente, pues muchas de ellas están enfocadas en solucionar un problema concreto del dominio en general (por ejemplo, bases de datos sobre el tratamiento de proteínas) y en este trabajo se estudian con el objetivo de cubrir las necesidades del MCGH. La perspectiva holística que proporciona el MCGH que se presentará posteriormente, proporciona una vista conceptual global para todas esas vistas parciales que cada base de datos aborda en forma particular. Esa es una de las contribuciones más relevantes de esta Tesis, y por ello es necesario revisar ese conjunto amplio y heterogéneo de bases de datos genómicas que abordan esas vistas parciales de datos. Para finalizar, la sección 3.3 expone las conclusiones sobre el estado del arte presentado/realizado.

3.1 Modelado Conceptual en el Dominio Genómico

En el campo de los *Sistemas de Información*, se utiliza el nombre de "*Modelado Conceptual*" para la actividad de obtener y describir el conocimiento general que se necesita saber para un sistema de información particular [1].

El objetivo principal del *modelado conceptual* es obtener esa descripción, la cual se denomina "*Esquema Conceptual*". Los *esquemas conceptuales* están escritos en lenguajes llamados "*Lenguajes de Modelado Conceptual*". El modelado conceptual es una parte importante de la ingeniería de requisitos, la primera y más importante fase en el desarrollo de un sistema de información [1], [44].

La aplicación de este enfoque en trabajos anteriores ha mostrado cómo los *modelos conceptuales* permiten proporcionar una definición clara del dominio, lo que permite que se puedan entender mucho mejor las entidades involucradas y sus relaciones. Por eso es ampliamente aceptado que la aplicación de modelado conceptual facilita la comprensión de dominios complejos, como, por ejemplo, el *genómico*.

Uno de los primeros trabajos presentados sobre la aplicación del modelado conceptual en el genoma, fue el de *Paton* [45], [46]. Este trabajo se enfocaba en describir el genoma desde distintas perspectivas, entre las que se encuentran la descripción del genoma de la célula eucariota, la interacción entre proteínas, el transcriptoma y otros componentes genéticos, sin embargo, su trabajo no tuvo una continuación fructífera en el dominio.

Para el año 2004, se presentó un trabajo en el cual se aplicaron principios de modelado conceptual en el contexto de las proteínas, las cuales incluían la consulta de grandes cantidades de datos y una estructura muy compleja. Este estudio se basó en la comparación y búsqueda en la estructura de una proteína en 3D, y este objetivo fue más fácil de alcanzar gracias a la aplicación del modelado conceptual [47].

También han surgido otras aproximaciones con el objetivo de representar los conceptos relacionados con el genoma, como, por ejemplo, la representación proporcionada por *GeneOntology*[4]. Este trabajo propone la unificación de términos –*surgió con la iniciativa de estandarizar la representación de los genes y sus atributos*-, para lo cual contaron con expertos en el campo de las ontologías de dominio, cuyo objetivo es caracterizar los términos usados en el dominio analizado (a modo de un "*thesaurus*" unificados de términos).

GeneOntology se presenta como un framework para el modelado de la biología. En esta se definen los conceptos y clases utilizados para describir la función de los genes, y las relaciones entre estos conceptos. Clasifica las funciones con respecto a tres puntos:

1) *función molecular*: actividades moleculares de los productos génicos;
2) *componente molecular*: donde los productos génicos están activos; y
3) *proceso biológico*: "*pathways*" y procesos más grandes compuestos por las actividades de múltiples productos génicos [48], [49].

Esta solución está orientada a la solución de un problema muy específico, sin embargo, en este trabajo se busca solucionar un problema global que repercute en el contexto bioinformático.

Hoy en día, a pesar de la gran cantidad de repositorios de datos genómicos disponibles públicamente, no es habitual encontrar modelos conceptuales estables -*subyacentes*-. Esto se debe principalmente a que la mayoría de los ficheros accesibles se sitúan en el espacio de la solución, no abordan el proceso de conceptualización del dominio analizado, y requieren la integración continua de mejoras directamente en temas de almacenamiento.

[4] *http://www.geneontology.org/*

El modelado conceptual no sólo se emplea como un enfoque para describir y representar un dominio específico, sino que también ayuda en la producción de software. En particular, el enfoque MDD (*Model-driven development*) [50] ya se ha utilizado en el dominio bioinformático. Por ejemplo, este método fue aplicado en el trabajo de *Garwood* et. al. (2006) en el que se crearon interfaces de usuario para consultar repositorios de datos biológicos [51].

Uno de los beneficios esenciales del modelado conceptual es que permite representar con precisión los conceptos relevantes del dominio analizado. En este trabajo se ha utilizado este enfoque para definir un modelo conceptual que representa las características y el comportamiento del genoma humano. En los capítulos 4 y 5 de este trabajo se presenta la evolución conceptual del modelo y la propuesta de integración y extensión del modelo respectivamente. Esto se debe a que en este dominio se requiere estar constantemente alineado con el nuevo conocimiento genómico adquirido.

En un contexto como el genómico en el que el conocimiento está en constante cambio, la evolución natural del *Modelo Conceptual del Genoma Humano* (MCGH) es fundamental, pues de esta manera se integran los nuevos descubrimientos en el dominio con el objetivo de mejorar el procesamiento de los datos genómicos, que tienen además que cumplir con el objetivo de contribuir y garantizar una *medicina de precisión (o medicina personalizada)* eficaz y real.

3.2 Bases de Datos Genómicas

La comunidad médica e informática ha realizado una gran cantidad de estudios con el objetivo de encontrar soluciones convincentes a los problemas de gestión para las bases de datos genómicas. Dichos inconvenientes radican en la necesidad de gestionar grandes fuentes de datos, por lo que se requiere mayor inversión en tiempo, almacenamiento e investigación, entre otros aspectos que exigen una mejora continua.

M. Rouse (2014) define el término de *base de datos* como "*una colección de información organizada de tal modo que sea fácilmente accesible, gestionada y actualizada*" [52], [53].

Como se ha comentado anteriormente, el dominio genómico es un entorno que supone un gran desafío con su objetivo final de "*entender y manipular el genoma*". Hoy en día existen un gran número de bases de datos biológicas. De acuerdo con el trabajo publicado sobre el catálogo de bases de datos del NAR[5] (*Nucleic Acids Research*), el año 2016 actualizaron su catálogo y se añadieron 88 repositorios (*recursos*) nuevos, eliminando un total de 23 sitios web obsoletos y dejando finalmente un listado de hasta 1,685 bases de datos biológicas [54].

La HGVS[6] (*Human Genome Variation Society*) también ofrece un catálogo de bases de datos genómicas (por ejemplo, variaciones/mutaciones), para el año 2015 contaban con un total de 1,755 repositorios listados [55].

De este extenso número de repositorios, la gran mayoría están centrados/enfocados en la solución de un aspecto específico de todo el genoma. En este dominio se presentan múltiples casos que dificultan el tratamiento de los datos, dando a lugar a problemas de: *dispersión*, *heterogeneidad*, *redundancia*, *inconsistencia*, … En el momento de gestionar dichos datos, a esta situación se le denota de forma genérica como "*caos de datos genómicos*".

Con las herramientas y motores de búsqueda actuales se pueden resolver ciertos problemas, pero es fundamental definir un marco ontológico holístico que permita delimitar el conocimiento relevante y que posibilite crear una estructura clara y sencilla para la gestión efectiva y eficiente de los datos genómicos. Se debe potenciar el desarrollo de bases de datos con información delimitada y revisada ("*curated*"), para lograr una optimización del rendimiento para reportar diagnósticos de mayor precisión y calidad.

Para resolver este problema, en este trabajo de Tesis se presenta el desarrollo de la Base de Datos del Genoma Humano (HGDB), la cual está basada en el MCGH que se describe en el Capítulo 4. La HGDB se ha desarrollado con el objetivo de integrar todo el conocimiento existente en el dominio genómico basándose en una representación conceptual holística del dominio. Para la carga de sus datos se realiza una serie de estudios y análisis que permitan filtrar/seleccionar los datos relevantes. De esta forma se genera un repositorio con datos

[5] *http://www.oxfordjournals.org/nar/database/c/*

[6] *http://www.hgvs.org/content/databases-tools*

"*selectivos*", que facilitan la obtención de diagnósticos genómicos con soporte científico reciente, creíble y relevante.

Para entender la necesidad de disponer de ese repositorio conceptual universal de datos genómicos, es necesario entender la complejidad que introduce la diversidad existente en las fuentes de datos genómicos actuales. Con ese objetivo se presenta a continuación una breve descripción de algunas de las bases de datos genómicas más conocidas y utilizadas por los expertos del dominio (*genetistas*, *laboratorios*, *grupos hospitalarios*).

3.2.1 1000 Genomas

El proyecto 1000 Genomas[7], tiene como objetivo principal catalogar la variación genética en la especie humana, para lo que se analizó el genoma completo de 2504 personas de 26 poblaciones diferentes (Figura 9). Este proyecto ha permitido determinar las propiedades y la distribución de las variaciones raras y comunes, proporcionando información de los procesos que dan lugar a la diversidad genética, y permitiendo una mejor comprensión de las enfermedades [56], [57].

El proyecto también ha contribuido a caracterizar la historia y la demografía de las poblaciones humanas ancestrales, definiendo un origen común del ser humano hace entre 150.000 y 200.000 años en África, lo que propicia que las personas de ascendencia africana tiendan a tener niveles más altos de diversidad genética. A partir de este momento, grupos relativamente pequeños de seres humanos comenzaron a emigrar de África, para extenderse por todo el mundo, pero al ser un número de personas pequeño portaron solo una fracción de las variantes genéticas existentes. Este suceso propició un cuello de botella más notable en las poblaciones no africanas, en cuanto a variación genética. Los científicos, además, estimaron que un genoma típico difiere del genoma humano de referencia en 4,1 - 5 millones de variantes, siendo un total de 88 millones de variantes las que se han caracterizado [58], [59].

[7] *http://www.internationalgenome.org/home*

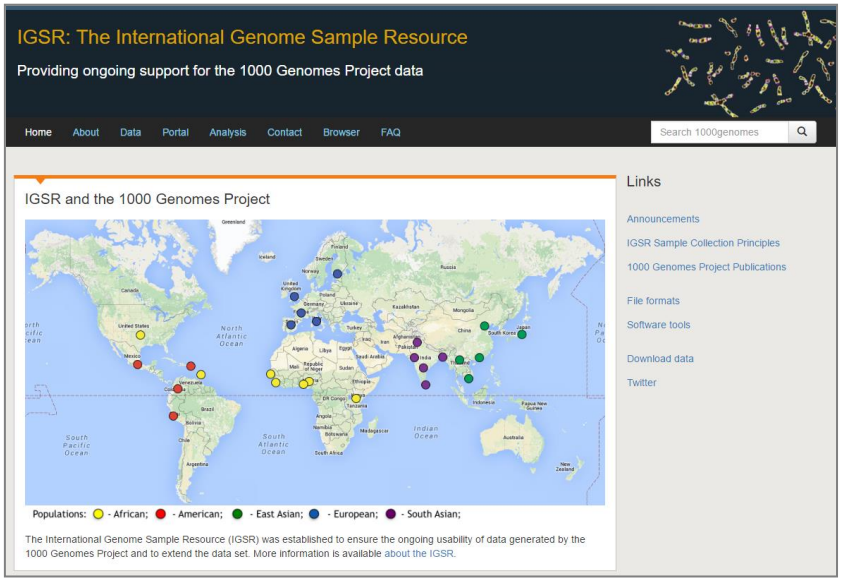


Figura 9. Proyecto 1000 Genomas (*website*)

Por un lado, más del 99% de las variantes genéticas consisten en SNPs (del inglés "*Single Nucleotide Polymorphisms*") y pequeñas inserciones y deleciones, siendo las variaciones estructurales las menos abundantes. Por otro lado, aunque las variantes comunes son compartidas alrededor del mundo, las variantes raras las encontramos generalmente restringidas a poblaciones estrechamente relacionadas. Este estudio tiene importantes implicaciones, pudiendo usarse para mejorar la comprensión y poder distinguir mejor entre la variación genética patógena y la no patógena lo que permite determinar factores de riesgo de enfermedades, entre otros aspectos [56].

3.2.2 ALFRED

La base de datos ALFRED (*the ALelle FREquency Database*) está diseñada para almacenar y diseminar frecuencias de alelos en sitios polimórficos humanos para múltiples poblaciones, enfocada principalmente para las comunidades de genética de poblaciones y antropología molecular [60], [61].

Actualmente, ALFRED posee información sobre 672 sitios polimórficos tipados en al menos una muestra de población y 288 poblaciones tipificadas para al menos un polimorfismo.

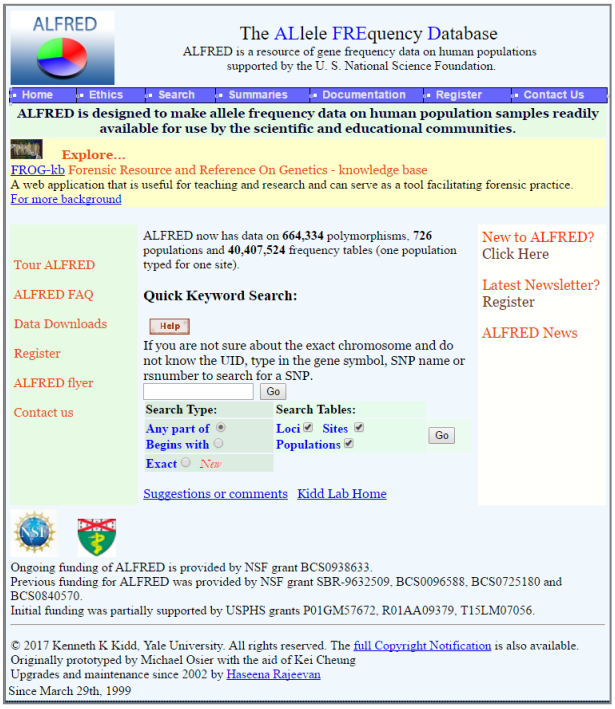


Figura 10. ALFRED (*website*)

Desde su lanzamiento inicial de la base de datos se han centrado en el aumento de la cantidad y calidad de los datos, haciendo enlaces recíprocos entre ALFRED y otras bases de datos relacionadas, y facilitando herramientas útiles para hacer los datos más comprensibles para el usuario final (por ejemplo, se proporcionan enlaces a las bases de datos moleculares para localizaciones de los polimorfismos y bases de datos antropológicas de información lingüística, etnográfica y demográfica sobre las poblaciones muestreadas. Las referencias a publicaciones están asociadas con las frecuencias y vinculadas a PubMed, siempre que sea posible) [62]. Este repositorio puede ser accedido directamente desde el sitio web de ALFRED (*http://alfred.med.yale.edu/alfred/index.asp*, ver Figura 10) [60].

3.2.3 BIC (Breast Cancer Information Core)

BIC[8] consiste en un repositorio central de información sobre mutaciones y polimorfismos en los genes de susceptibilidad al cáncer de mama. Es un ejemplo pues de fuente de datos genómicos con un propósito muy delimitado en el ámbito de un fenotipo concreto particular.

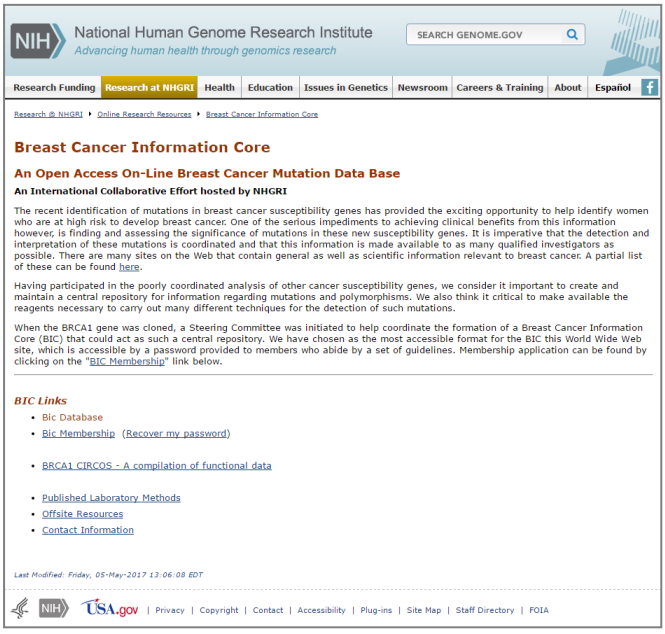


Figura 11. BIC (*website*)

El "*Breast Cancer Information Core*" (BIC) es una base de datos de mutación en línea de acceso abierto (Figura 11) para los genes de susceptibilidad al cáncer de mama, además de crear un catálogo de todas las mutaciones y polimorfismos en los genes de susceptibilidad al cáncer de mama.

Un objetivo principal de BIC es facilitar la detección y caracterización de estos genes, proporcionando apoyo técnico en forma de protocolos de detección de mutaciones, secuencias de cebadores[9] y reactivos de

[8] *https://research.nhgri.nih.gov/bic/*

[9] Cebadores: es la secuencia de inicio en la replicación de la cadena (sinónimos: *iniciador*, *primer*, etc.)

acceso. Este sitio posee información adicional en la cual incluye una revisión de la literatura compilada a partir de estudios publicados, enlaces a otros recursos de investigación e información sobre el cáncer de mama en internet. También se facilita un foro interactivo de discusión, el cual permite a los investigadores publicar o responder a preguntas y/o comentarios en un tablón especifico [63].

3.2.4 BioQ

BioQ proporciona herramientas de consulta y documentación para bases de datos relacionales *-genómicas-*. Las herramientas de documentación proporcionan información detallada sobre varias capas, incluyendo métodos para rastrear el flujo de procesos experimentales y determinar la fuente experimental de los datos (*http://bioq.saclab.net/*, ver Figura 12).

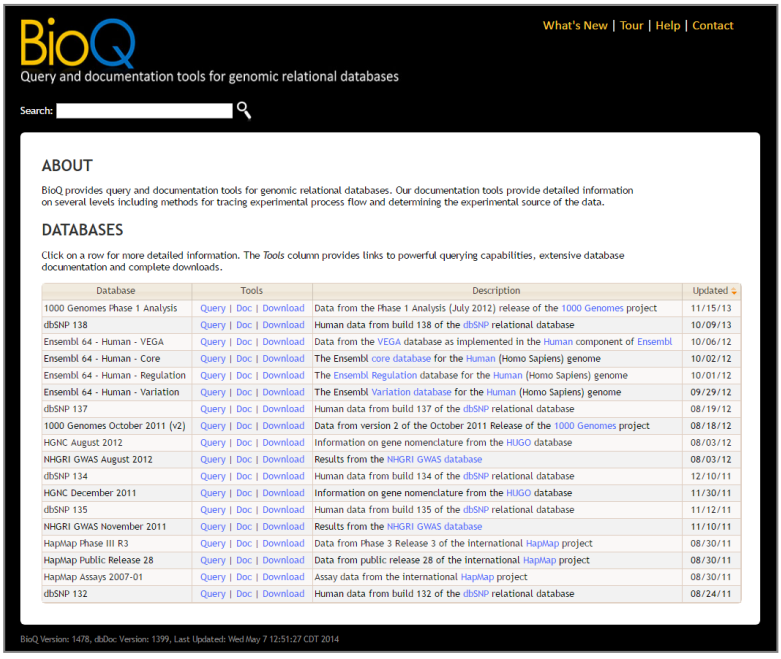


Figura 12. BioQ (*website*)

La motivación de esta herramienta es determinar sistemáticamente los orígenes experimentales de los datos (trazando sistemáticamente sus

orígenes experimentales a los sujetos originales y biológicos), lo que permite evaluar la fiabilidad de los datos.

BioQ permite a los investigadores tanto visualizar la procedencia de los datos como explorar elementos individuales de flujo de proceso experimental utilizando herramientas precisas para la exploración detallada de datos y documentación.

Esta herramienta incluye una serie de bases de datos de variación genética (*humana*), como, por ejemplo, los proyectos de *HapMap* y *1000 genomas*. BioQ se encuentra disponible de forma gratuita para el público en general desde su sitio web (Figura 12) [64].

3.2.5 ClinVar

La base de datos de ClinVar tiene un amplio alcance e incluye interpretaciones de variantes en cualquier región del genoma humano, incluyendo las mitocondrias[10]. Las variantes en ClinVar pueden ser de cualquier longitud o tipo, desde sustituciones de un solo nucleótido y pequeñas inserciones/deleciones hasta cambios de número de copias y reordenamientos citogenéticos. Estas variantes pueden haber sido identificadas en la línea germinal o en fuentes somáticas.

En general, las variantes de ClinVar se han observado en individuos y familias, mediante una investigación o en un entorno clínico, e interpretadas por su importancia clínica en relación con uno o más trastornos o con un conjunto de características clínicas y el modo de herencia [65], [66].

Algunas interpretaciones orientadas a la investigación pueden proporcionar significación funcional basada en la evidencia experimental, la cual puede informar la interpretación clínica de una variante por otros. ClinVar tiene actualmente más de 158,000 interpretaciones presentadas, representando más de 125,000 variantes. Las interpretaciones en la base de datos afectan a más de 26,000 genes, incluyendo variantes estructurales que pueden incluir muchos genes;

[10] Las mitocondrias son un tipo de orgánulos localizados en las células, que se encargan de suministrar la mayoría de la energía que se necesita en la actividad o respiración celular. Las mitocondrias funcionan como centrales energéticas en la célula, sintetizando ATP con los carburantes metabólicos, tales como: la glucosa, los ácidos grados y los aminoácidos (*http://funcionde.com/mitocondrias/*).

para las variantes que afectan a un solo gen, casi 4,800 genes están representados en ClinVar [66]. ClinVar agrega información sobre la variación genómica y su relación con la salud humana (*https://www.ncbi.nlm.nih.gov/clinvar/*, ver Figura E).

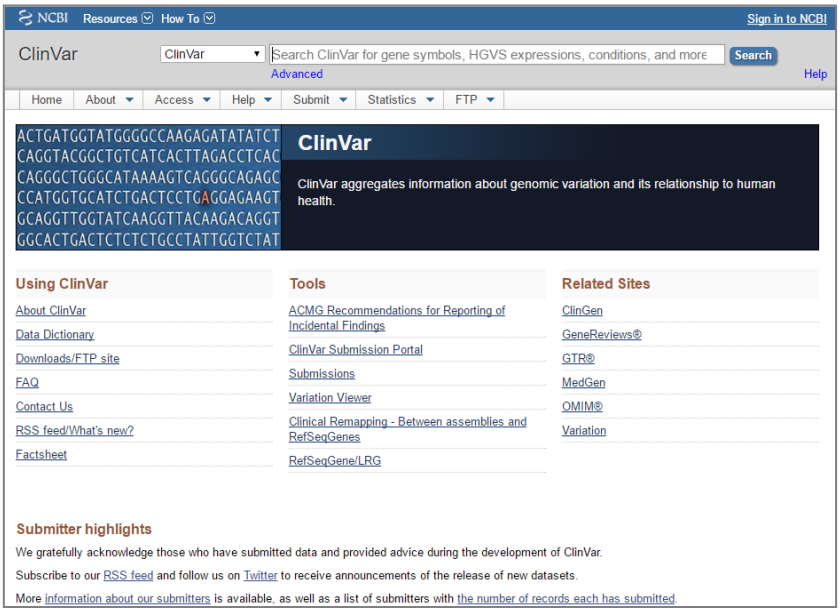


Figura 13. ClinVar (*website*)

3.2.6 COSMIC (Catalogue of somatic mutations in cancer)

COSMIC es un sistema de base de datos diseñado para reunir información a nivel mundial sobre las mutaciones somáticas[11] en el cáncer humano en un único sistema y plantearlo de una manera fácilmente explorable [67], [68]. Este repositorio está compuesto por la recopilación de datos asociados a *publicaciones científicas* y *estudios experimentales* a gran escala del *Proyecto Genoma del Cáncer* (en inglés, *Cancer Genome Project*), llevado a cabo por el *Instituto Sanger de*

[11] *Mutaciones somáticas*: alteración del ADN que ocurre después de la concepción. Las mutaciones somáticas se pueden presentar en cualquiera de las células del cuerpo, excepto las células germinativas (esperma y huevo) y, por lo tanto, no pasan a los hijos. Estas alteraciones pueden causar cáncer u otras enfermedades (pero esto no siempre ocurre) [197].

Cambridge. COSMIC está disponibles a través de su sitio web -*gratuita y sin restricciones*- (*http://cancer.sanger.ac.uk/cosmic*, ver Figura 14).

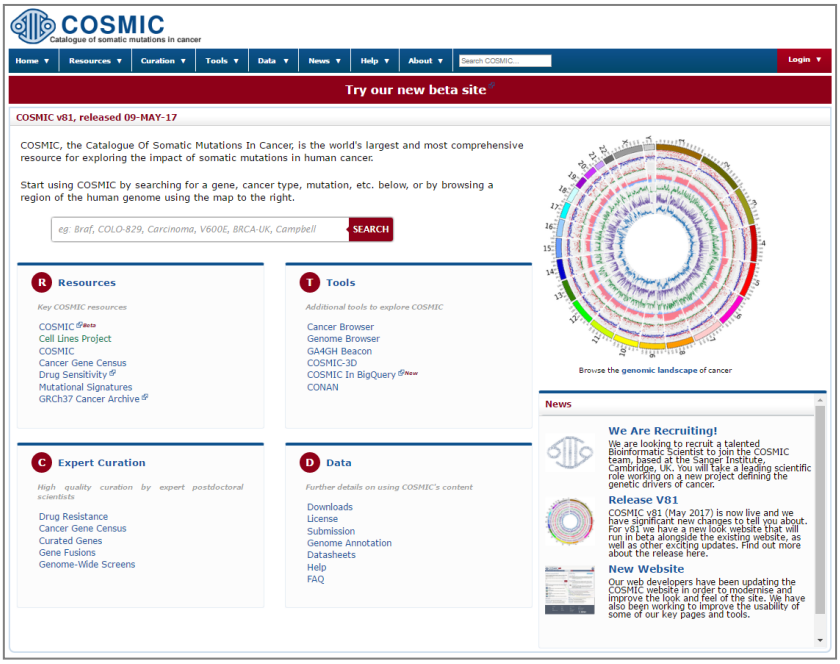


Figura 14. COSMIC (*website*)

Esta base de datos fue lanzada en 2004, y en ese entonces contaba con datos de únicamente 4 genes: "*HRAS*", "*KRAS2*", "*NRAS*" y "*BRAF*". En los últimos 10 años han visto un gran crecimiento en las áreas de *genética del cáncer* y la *genómica*, por lo que actualmente COSMIC cuenta con 136 genes curados y 12,542 genomas de cáncer (ver más detalles en la Tabla 4).

Tabla 4. Contenido total en la versión 70 de la base de datos COSMIC (*versión de agosto-2014*), tabla extraída de [69].

Genes (transcripts)	28 735
Tumor samples	1 029 547
Coding mutations	2 002 811
Curated publications	19 703
Fusion mutations	10 435
Genomic rearrangements	61 299
Whole genomes	12 542
Copy number aberrations	695 504
Gene expression variants	60 119 787

Este proyecto se creó originalmente para detallar mutaciones genéticas de genes codificantes (simples), pero al día de hoy describe millones de mutaciones codificantes, mutaciones no codificantes, reordenamientos genómicos, fusión de genes, anomalías en el número de copias y variantes de expresión génica en todo el genoma humano [69].

3.2.7 dbGAP

La base de datos de *genotipos* y *fenotipos* (dbGap, *http://www.ncbi.nlm.nih.gov/gap*) es un repositorio patrocinado por los Institutos Nacionales de Salud con el objetivo de *archivar*, *curar* y *distribuir* información producida por estudios que investigan la interacción entre el genotipo y fenotipo (Figura 15).

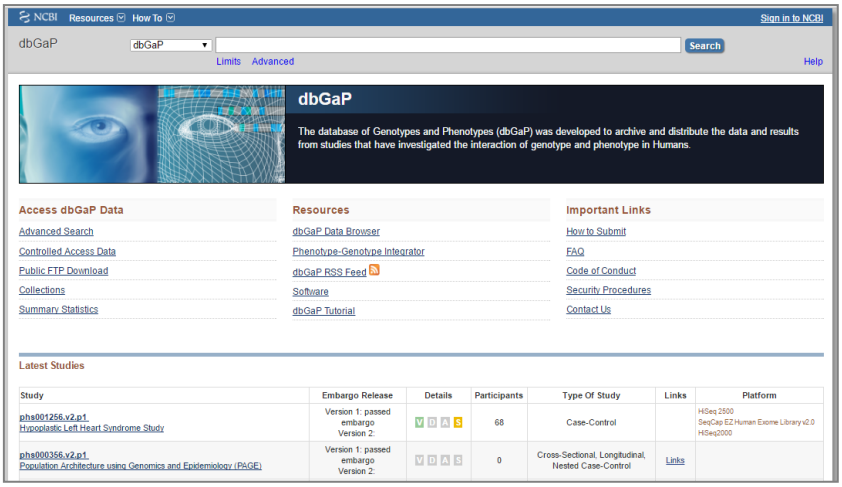


Figura 15. dbGAP (*website*)

La información en dbGAP está organizada como una estructura jerárquica que incluye los siguientes objetos: fenotipos (como variables y conjuntos de datos), diversos datos de análisis molecular (SNP y "*Expression Array Data*", secuencia y marcas epigenómicas), análisis y documentos (por ejemplo, resultados de análisis, imágenes médicas, información general sobre el estudio, documentos que contextualizan variables fenotípicas –*protocolos de investigación y cuestionarios*-).

Los metadatos accesibles al público sobre los estudios presentados, los resúmenes de los niveles de los datos y los documentos relacionados con

los estudios se pueden consultar libremente en el sitio web de dbGAP. En el caso de los datos de nivel individual son accesibles a través de la aplicación de acceso controlado por científicos de todo el mundo [70].

3.2.8 dbSNP

La "*Single Nucleotide Polymorphism database*" (dbSNP[12]) es un archivo de dominio público, el cual cuenta con una amplia colección de polimorfismos genéticos simples. Desde su creación en septiembre de 1998, la base de datos "*dbSNP*" ha servido como un repositorio central y público para el tema de *variaciones genéticas* (Figura 16) [71].

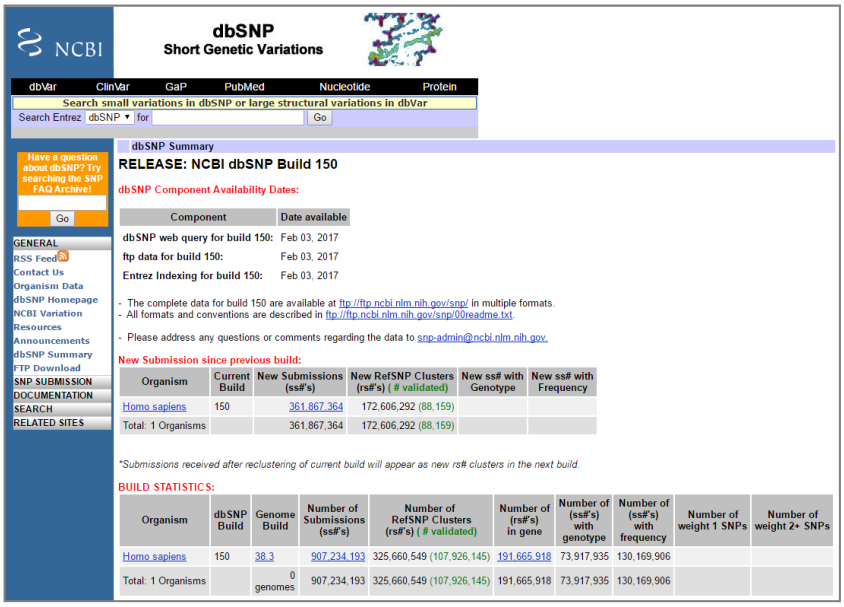


Figura 16. dbSNP (*website*)

Actualmente, dbSNP clasifica las variaciones de la secuencia de nucleótidos de la siguiente forma (y composición porcentual de la base de datos): a) *sustituciones de un solo nucleótido*, 99.77%; b) *polimorfismos con inserción / deleción pequeña*, 0.21%; c) *secuencia con regiones invariantes*, 0.02%; d) *repeticiones de microsatélite*, 0.001%; e) *variantes nombradas*, < 0.001%; y f) *ensayos heterocigotos no*

[12] *https://www.ncbi.nlm.nih.gov/projects/SNP/*

caracterizados, < 0.001%. No hay ningún requisito o suposición sobre las frecuencias alélicas mínimas o neutralidad funcional para los polimorfismos en la base de datos, por lo que, el alcance de dbSNP incluye la enfermedad que causa la mutación clínica, así como polimorfismos neutros. Además, de los identificadores de registro asignados tanto por los solicitantes como por el NCBI, las entradas realizadas en dbSNP registran la información de la secuencia alrededor del polimorfismo, las condiciones experimentales especificas (las necesarias para realizar un experimento), descripciones de la población que contiene la variación e información de la frecuencia por población o genotipo individual. Aunque actualmente la mayoría de los trabajos subidos son para "*Homo sapiens*", la base de datos dbSNP tiene trabajos sobre el "*Mus musculus*", y en general esta base de datos puede aceptar información sobre variaciones de cualquier especie y parte de un genoma particular [71], [72].

dbSNP está actualmente integrada con otras grandes bases de datos públicas –*de variaciones*-, como, por ejemplo, la base de datos de *NCI CGAP-GAI of EST-derived SNPs* [73], la *TSC (The SNP Consortium, Ltd) variation initiative* [73] y *HGBASE* [74].

3.2.9 D-HaploDB (Definitive Haplotype Database)

La Base de Datos de Haplotipo Definitivo (D-HaploDB) es un recurso accesible en la Web de haplotipos definitivos del genoma, determinados a partir de una colección japonesa de "*moles hidatidiformes completos*" (CHMs, *complete hydatidiform moles*), cada uno de los cuales lleva un genoma derivado de un único espermatozoide [75].

Para el año 2014 la base de datos agrupaba genotipos para 281,439 SNPs comunes de 74 CHMs que se determinaron mediante una tecnología de hibridación de oligonucleótidos basados en array de alto rendimiento. La base de datos también presenta mapas de los bloques de haplotipos y de desequilibrio de enlace binarios junto con tagSNPs que podría resultar útil para los estudios de asociación de genes de la enfermedad [76].

La relación crítica entre las muestras en este estudio es poco probable, porque la formación de un CHM es un evento maternal de

rara aparición esporádica, y su genotipo es el de los espermatozoides entrantes. Esto se demuestra por la ausencia de haplotipos compartidos largos extendidos (ESHs). El recurso D-HaploDB es de libre acceso a través de Internet (*http://orca.gen.kyushu-u.ac.jp*) [75].

3.2.10 DisGeNET

DisGeNET es una plataforma de descubrimiento integral diseñada para abordar una variedad de preguntas relacionadas con el fundamento genético de las enfermedades humanas. La versión actual de DisGeNET[13] (v5.0) contiene 561,119 asociaciones de enfermedades genéticas (GDA), entre 17,074 genes y 20,370 enfermedades, trastornos, rasgos y fenotipos humanos *–clínicos o anormales-*, y 135,588 asociaciones de variantes-enfermedad (VDA), entre 83,002 SNPs y 9,169 enfermedades y fenotipos (Figura 17) [77].

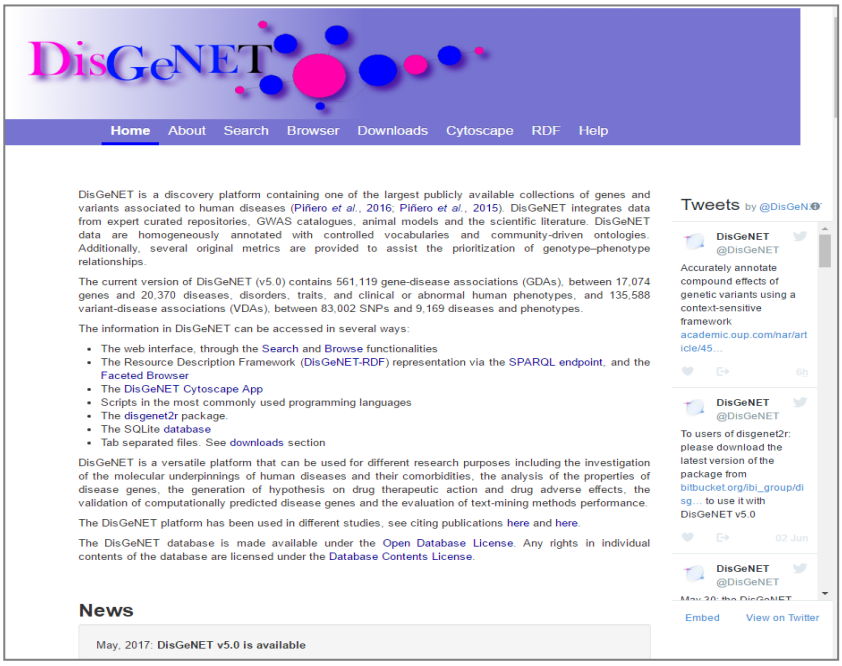


Figura 17. DisGeNET (*website*)

[13] *http://www.disgenet.org/web/DisGeNET/menu*

DisGeNET integra bases de datos "*curadas*" por expertos –*con datos extraídos de textos científicos*-, y cubre información sobre enfermedades mendelianas y complejas, e incluye datos de modelos de enfermedades animales. Además, cuenta con una puntuación basada en la evidencia de apoyo para priorizar las asociaciones entre genes y enfermedades [77].

DisGeNET es un recurso de acceso abierto que está disponible a través de una interfaz web, un complemento llamado "*Cytoscape*" y como un recurso de Web Semántica.

La interfaz web soporta la exploración y navegación de los datos de forma amigable para el usuario. Esta base de datos ofrece una de las colecciones más completas sobre asociaciones de genes y enfermedades humanas. Esto facilita un valioso conjunto de herramientas para investigar los mecanismos moleculares subyacentes a las enfermedades de origen genético, diseñadas para satisfacer las necesidades de diferentes perfiles de usuarios, incluyendo bioinformáticos y profesionales del cuidado de la salud [77].

3.2.11 Ensembl

Ensembl es un sistema de interpretación genómica. Se trata de un navegador de genomas de vertebrados que apoya la investigación en genómica comparativa, la evolución, la secuencia de variación y la regulación transcripcional.

Ensembl (*http://www.ensembl.org*) anota los genes, calcula múltiples alineaciones, predice la función reguladora y recoge los datos de la enfermedad (Figura 18). Las herramientas de Ensembl incluyen: *BLAST*, *BLAT*, *BioMart* y *VEP* (*Variant Effect Predictor*) para todas las especies soportadas [78]–[80].

Este proyecto es mantenido por el *EMBL*, el *EBI* y el *Welcome Trust Sanger Institute*. Es posible encontrar para un gen determinado su estructura intrónica-exónica, su localización en el genoma, sus variantes polimórficas ya sean en forma de SNP o splicing alternativos del gen [81].

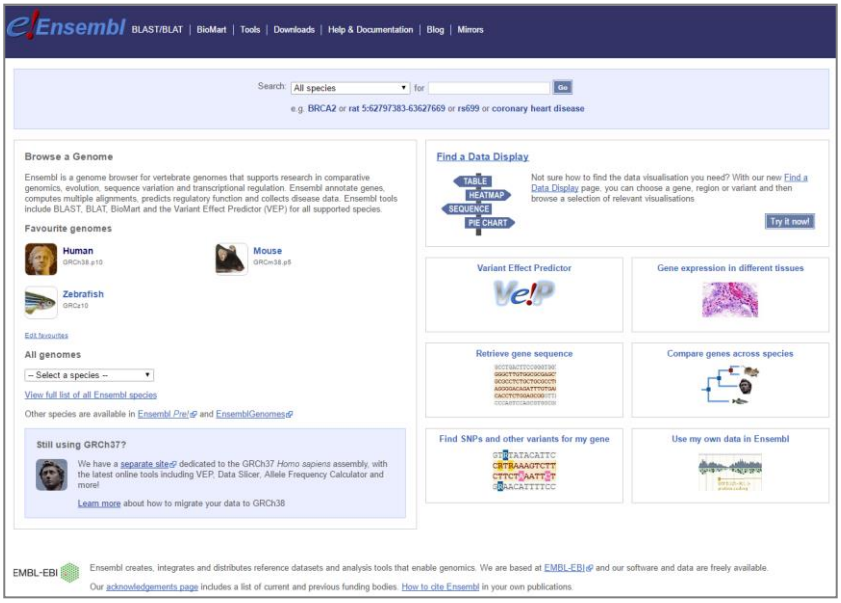
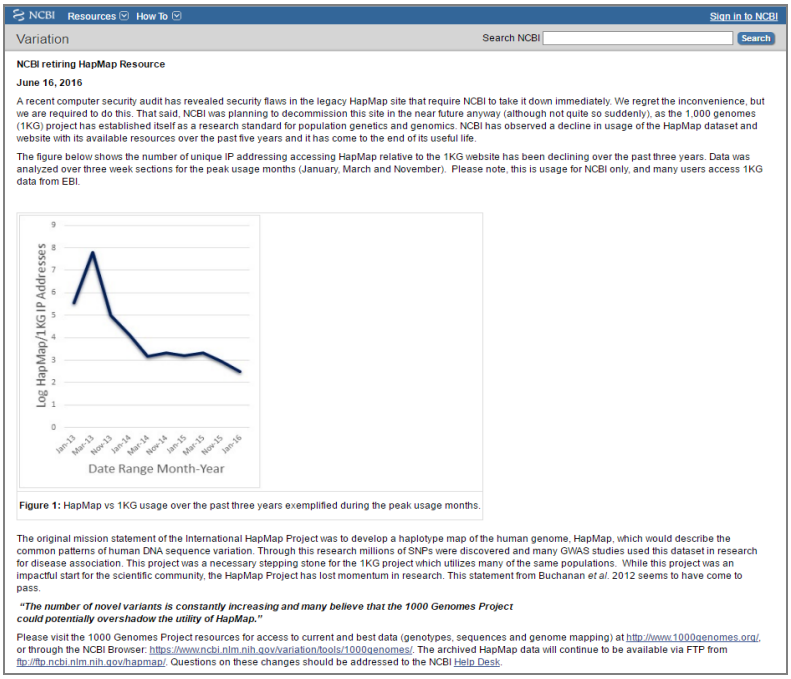


Figura 18. Ensembl (*website*)

3.2.12 HapMap

El objetivo del Proyecto Internacional HapMap consistía en crear un mapa de haplotipos del genoma humano (Figura 19). A menudo conocido como *HapMap*, describe los patrones comunes de la variación genética humana [82]. HapMap proporciona un recurso clave que los investigadores pueden usar para encontrar genes que afectan a la salud, la enfermedad y las respuestas a los medicamentos y los factores ambientales. La información producida por el proyecto está ahora disponible gratuitamente en bases de datos públicas a los investigadores alrededor del mundo [83]–[85].

El Proyecto Internacional HapMap comenzó oficialmente con una reunión, celebrada del 27-29 de octubre de 2002, y logró su objetivo de completar el mapa en un plazo de tres años. El proyecto fue una colaboración entre investigadores de centros académicos, grupos de investigación biomédica sin fines de lucro y empresas privadas en *Japón*, el *Reino Unido*, *Canadá*, *China*, *Nigeria* y los *Estados Unidos* (Figura 20). Puede encontrarse una lista de instituciones participantes y de financiación en: *http://hapmap.ncbi.nlm.nih.gov/groups.html*.



NCBI Resources How To Sign in to NCBI

Variation Search NCBI Search

NCBI retiring HapMap Resource

June 16, 2016

A recent computer security audit has revealed security flaws in the legacy HapMap site that require NCBI to take it down immediately. We regret the inconvenience, but we are required to do this. That said, NCBI was planning to decommission this site in the near future anyway (although not quite so suddenly), as the 1,000 genomes (1KG) project has established itself as a research standard for population genetics and genomics. NCBI has observed a decline in usage of the HapMap dataset and website with its available resources over the past five years and it has come to the end of its useful life.

The figure below shows the number of unique IP addressing accessing HapMap relative to the 1KG website has been declining over the past three years. Data was analyzed over three week sections for the peak usage months (January, March and November). Please note, this is usage for NCBI only, and many users access 1KG data from EBI.

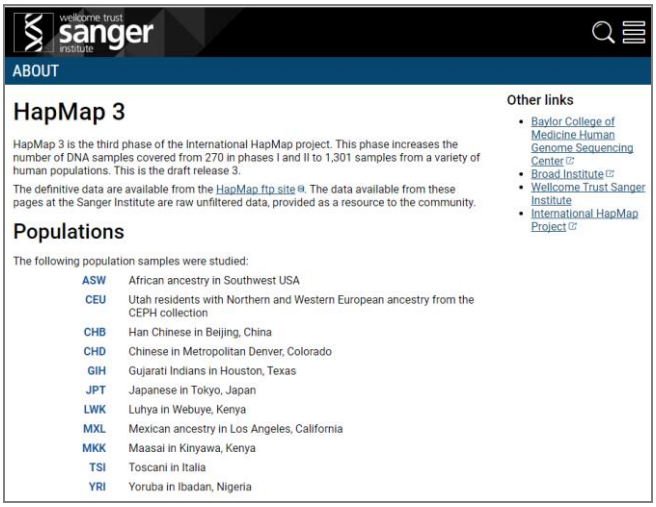
Figure 1: HapMap vs 1KG usage over the past three years exemplified during the peak usage months.

The original mission statement of the International HapMap Project was to develop a haplotype map of the human genome, HapMap, which would describe the common patterns of human DNA sequence variation. Through this research millions of SNPs were discovered and many GWAS studies used this dataset in research for disease association. This project was a necessary stepping stone for the 1KG project which utilizes many of the same populations. While this project was an impactful start for the scientific community, the HapMap Project has lost momentum in research. This statement from Buchanan *et al.* 2012 seems to have come to pass.

"The number of novel variants is constantly increasing and many believe that the 1000 Genomes Project could potentially overshadow the utility of HapMap."

Please visit the 1000 Genomes Project resources for access to current and best data (genotypes, sequences and genome mapping) at http://www.1000genomes.org/, or through the NCBI Browser: https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/. The archived HapMap data will continue to be available via FTP from ftp://ftp.ncbi.nlm.nih.gov/hapmap/. Questions on these changes should be addressed to the NCBI Help Desk.

Figura 19. HapMap (*website*)



wellcome trust sanger institute

ABOUT

HapMap 3

HapMap 3 is the third phase of the International HapMap project. This phase increases the number of DNA samples covered from 270 in phases I and II to 1,301 samples from a variety of human populations. This is the draft release 3.

The definitive data are available from the HapMap ftp site. The data available from these pages at the Sanger Institute are raw unfiltered data, provided as a resource to the community.

Populations

The following population samples were studied:

ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

Other links

- Baylor College of Medicine Human Genome Sequencing Center
- Broad Institute
- Wellcome Trust Sanger Institute
- International HapMap Project

Figura 20. Poblaciones tratadas Proyecto HapMap (*3era. Fase*)

3.2.13 HGMD

La Base de Datos de Mutaciones de Genes Humanos (*Human Gene Mutation Database:* HGMD®) consiste en una amplia colección de mutaciones de la línea germinal de genes nucleares asociados a enfermedades humanas hereditarias.

Esta base de datos fue desarrollada originalmente para el estudio de mecanismos mutacionales en genes humanos, aunque actualmente ha adquirido una utilidad mucho más amplia ya que representa una fuente de referencia actualizada del espectro de lesiones heredables en genes humanos [86].

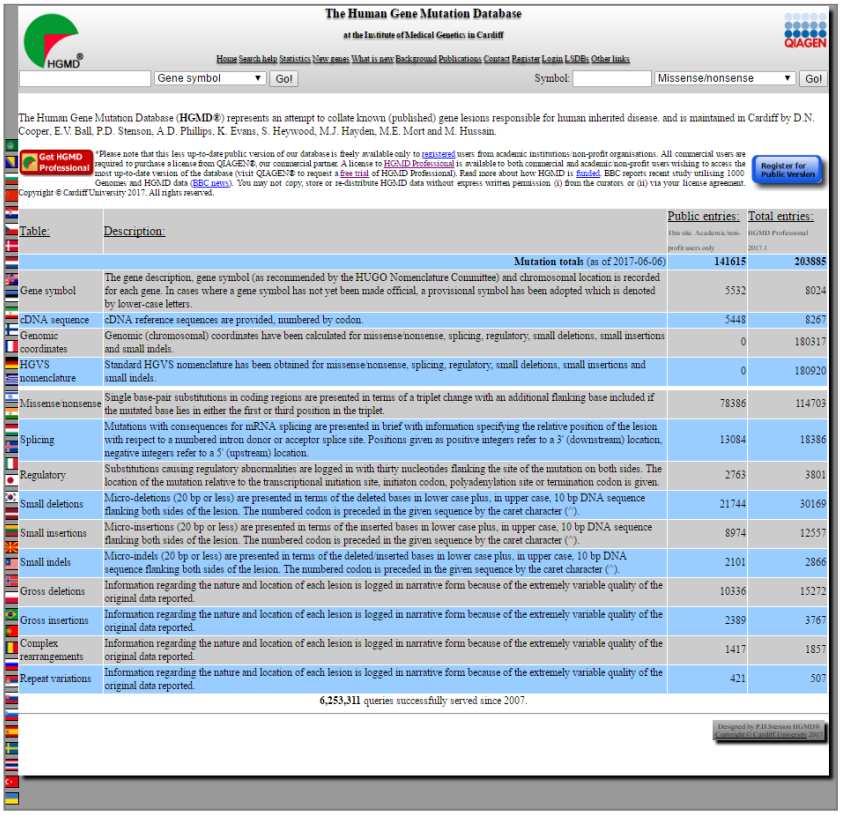


Figura 21. HGMD (*website*)

HGMD[14] posee todas las mutaciones de la línea germinal que producen enfermedades y los polimorfismos, tanto funcionales como los asociados a enfermedades, descritos en la literatura, y provee estos datos en un formato de fácil acceso para todos aquellos interesados, ya sean del ámbito académico, clínico o comercial [87].

Actualmente, HGMD constituye, de facto, la principal base de datos de mutaciones humanas asociadas a enfermedades disponible para la comunidad científica (Figura 21). Los datos proveen sustituciones de un único nucleótido en regiones codificantes (ejemplo, mutaciones sinónimas y no-sinónimas), regiones regulatorias y sitios relevantes de '*splicing*' en genes nucleares humanos, micro-deleciones, micro-inserciones, '*indels*', expansiones de elementos repetitivos, así como importantes lesiones génicas (*deleciones*, *inserciones* y *duplicaciones*) y re-arreglos complejos de genes.

HGMD es de acceso gratuito para los usuarios registrados, académicos o sin ánimo de lucro. Los datos de mutaciones están disponibles en este sitio web luego de 3 años de su inclusión inicial. La subscripción a la versión actualizada de HGMD (*HGMD Professional*), tanto para usuarios académicos como comerciales se encuentra bajo la licencia de sus socios comerciales, *BIOBASE GmbH*.

La versión profesional (HGMD Professional, la cual se actualiza cada 3 meses) facilita además herramientas avanzadas de búsqueda e información adicional específica de genes y mutaciones ausentes en la versión pública [88].

3.2.14 KEGG

KEGG (Enciclopedia de Genes y Genomas de Kioto, en inglés "*Kyoto Encyclopedia of Genes and Genomes*") es un recurso de base de datos para comprender funciones de alto nivel y utilidades del sistema biológico, como, la célula, el organismo y el ecosistema, a partir de información de nivel molecular, especialmente conjuntos de datos moleculares a gran escala generados mediante la secuenciación del genoma y otras tecnologías experimentales de alto rendimiento (*http://www.genome.jp/kegg/*) (Figura 22).

[14] *http://www.hgmd.cf.ac.uk/ac/index.php*

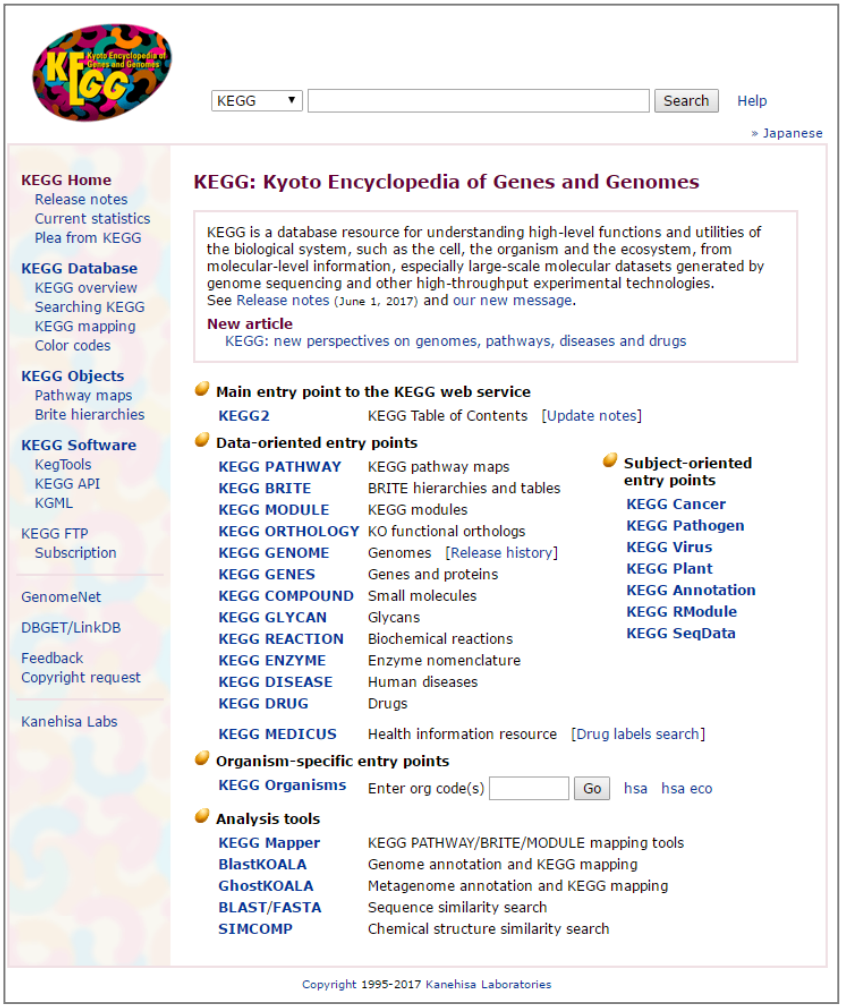


Figura 22. KEGG (*website*)

La información genómica se almacena en la base de datos "*GENES*", que es una colección de catálogos de genes para todos los genomas completamente secuenciados y algunos genomas parciales con la anotación actualizada de las funciones de los genes. La información funcional de orden superior se almacena en la base de datos "*PATHWAY*", que contiene representaciones gráficas de procesos

celulares, tales como metabolismo, transporte de membrana, transducción[15] de señales y ciclo celular.

La base de datos "*PATHWAY*" se complementa con un conjunto de tablas de grupos ortólogos para la información sobre las subpistas conservadas (motivos de la ruta –*pathway*-), que a menudo están codificadas por genes acoplados en posición en el cromosoma y que son especialmente útiles en la predicción de las funciones de los genes. Una tercera base de datos en KEGG es "*LIGAND*" para la información sobre compuestos químicos, moléculas de enzimas y reacciones enzimáticas.

KEGG proporciona herramientas de gráficos *Java* para la navegación en mapas del genoma, la comparación de dos mapas del genoma y la manipulación de mapas de expresión, así como herramientas computacionales para la comparación de secuencias, la comparación de gráficos y el cálculo de trayectoria. Las bases de datos KEGG se actualizan diariamente y se ponen a disposición de forma gratuita [89].

3.2.15 LOVD (Leiden Open-source Variation Database)

Las *Locus-Specific DataBases* (LSDBs) almacenan información sobre la variación de la secuencia génica asociada con fenotipos humanos y son utilizadas con frecuencia como referencia por investigadores y clínicos (Figura 23).

La base de datos LOVD (*Leiden Open-source Variation Database*) fue desarrollada como un paquete *LSDB-in-a-Box* basada en la web e independiente de la plataforma. LOVD fue diseñada con el objetivo de ser fácil de configurar y mantener, además de que sigue las recomendaciones de la *Sociedad de Variantes del Genoma Humano* (HGVS) [90] (*http://www.lovd.nl/3.0/home*).

En el trabajo de *Fokkema* et. al. (2011) se describe LOVD en su versión 2.0, la cual incorpora mayor flexibilidad, funcionalidad y tiene la capacidad de almacenar variantes de secuencias en múltiples genes por paciente.

[15] *Transducción*: por definición, es la transformación de un tipo de señal o energía en otra de distinta naturaleza.

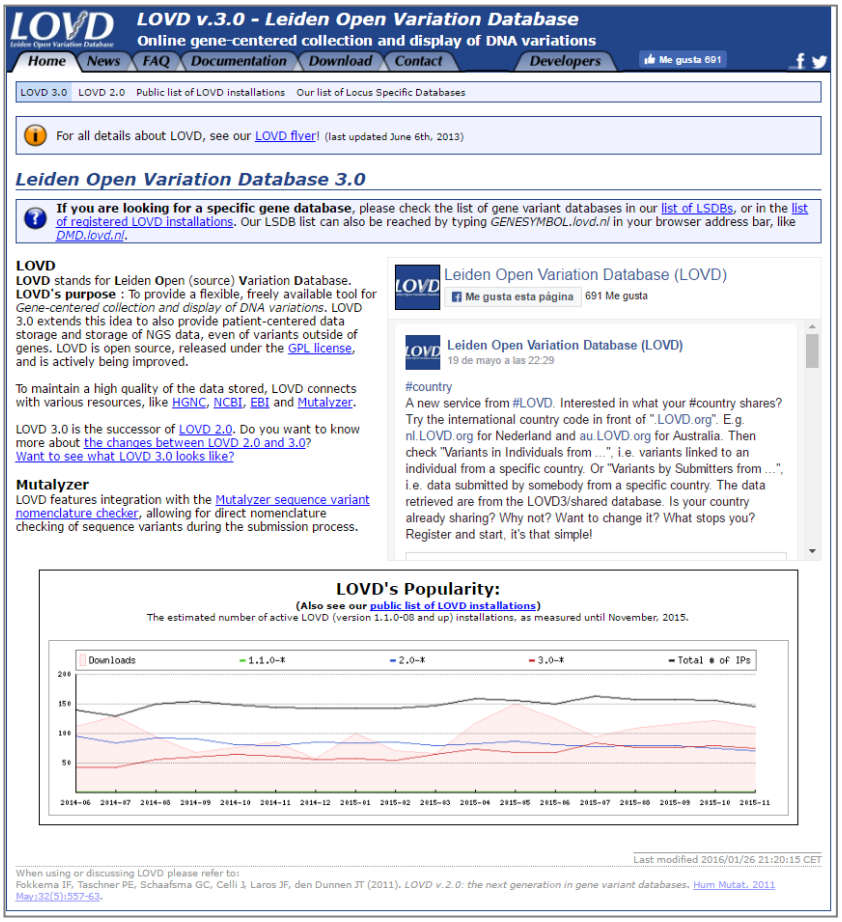


Figura 23. LOVD (*website*), para consultar la guía de usuario de la versión 3.0 ver el trabajo [91].

Para reducir la redundancia, el paciente y los datos de variantes de secuencia se almacenan en tablas separadas. Dichas tablas están vinculadas para generar conexiones entre los datos de la secuencia variante para cada gen y cada paciente.

La estructura dinámica permite a los gestores de bases de datos añadir columnas personalizadas. La estructura de la base de datos soporta consultas rápidas y permite el almacenamiento de variantes de secuencias a partir del análisis de secuencias de alto rendimiento.

LOVD contiene medidas para garantizar la seguridad de la misma (incluye –*el acceso no autorizado*-). En la actualidad, el sitio web de LOVD lista 71 instalaciones públicas de la base de datos, las cuales alojan 3,294 bases de datos de variantes genéticas con 199,000 variantes en 84,000 pacientes. Con el fin de promover la estandarización LSDB, y por lo tanto la interoperabilidad de la base de datos, LOVD ofrece a sus usuarios un espacio libre en su servidor y ayuda a establecer un LSBD en el servidor Leiden [90].

3.2.16 OMIM (Online Mendelian Inheritance in Man™)

La base de datos OMIM es una base de datos de conocimiento fácil de entender, de información contrastada y útil (según defienden sus autores) de genes humanos y desordenes genéticos compilados para apoyar la investigación y educación sobre genética humana y la práctica de la genética clínica.

Este proyecto fue iniciado por el Dr. Victor A. McKusick como la referencia definitiva de herencia mendeliana en el hombre. OMIM (*https://www.omim.org/*) se deriva de la literatura biomédica, y está escrito y editado por la *Universidad Johns Hopkins* con la participación de científicos y médicos de todo el mundo (Figura 24) [92], [93].

Cada entrada OMIM contiene un resumen completo de un fenotipo y/o gen determinado genéticamente. Éste tiene numerosos enlaces a otras bases de datos genéticas como: secuencias de ADN y proteínas, referencias de PubMed, bases de datos de mutaciones tanto generales/genéricas como de locus-específicas, nomenclatura HUGO, grupos de apoyo al paciente, entre otros más.

OMIM es un portal fácil y directo para la creciente información en genética humana [94], [95].

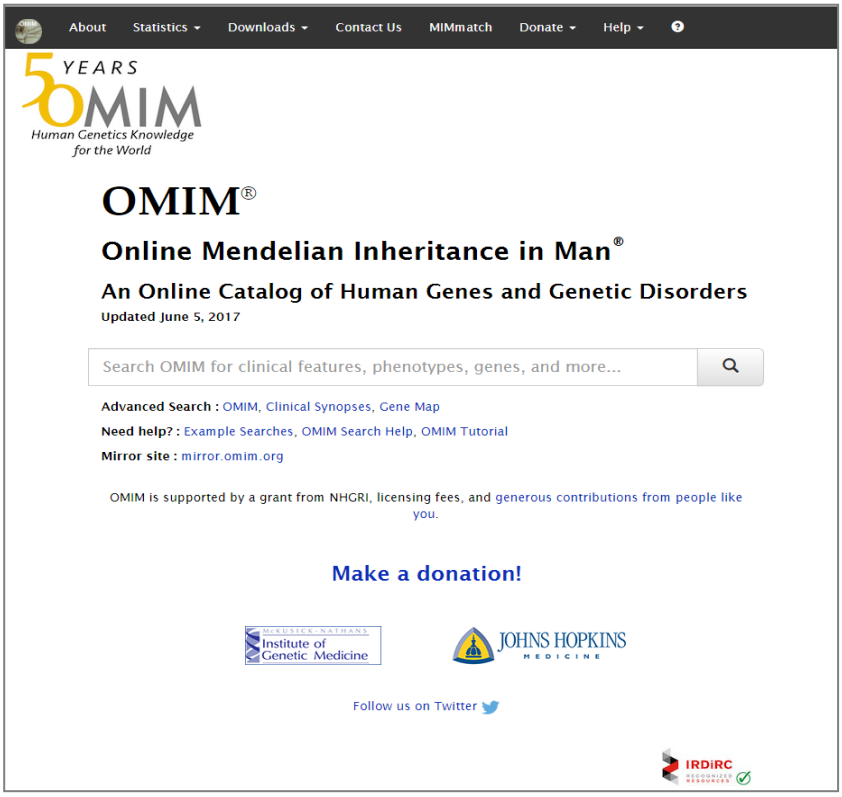


Figura 24. OMIM (*website*)

3.2.17 REACTOME

La base de conocimientos REACTOME (*www.reactome.org*) proporciona detalles moleculares de la traducción de señales, el transporte, la replicación del ADN, el metabolismo y otros procesos celulares como una red ordenada de transformaciones moleculares –*una versión ampliada de un mapa metabólico clásico, en un único modelo de datos coherente*- (Figura 25) [96].

REACTOME funciona tanto como un archivo de procesos biológicos como una herramienta para descubrir relaciones funcionales inesperadas en datos, tales como estudios de patrones de expresión génica o catálogos de mutaciones somáticas de células tumorales.

Figura 25. REACTOME (*website*)

Durante los dos últimos años se han re-desarrollado los componentes principales de la interfaz web de REACTOME, para de esta forma mejorar la usabilidad, la capacidad de respuesta y visualización de los datos [97], [98]. *Rodríguez-Tarduchy* explica que "*la idea de REACTOME sería representar los diferentes protagonistas de un determinado proceso biológico estableciendo relaciones e interacciones entre los mismos*" [99].

3.2.18 SNPedia

SNPedia (*http://www.SNPedia.com*) es un recurso wiki sobre las consecuencias funcionales de la variación genética humana, tal como se publicó en estudios revisados por pares (Figura 26).

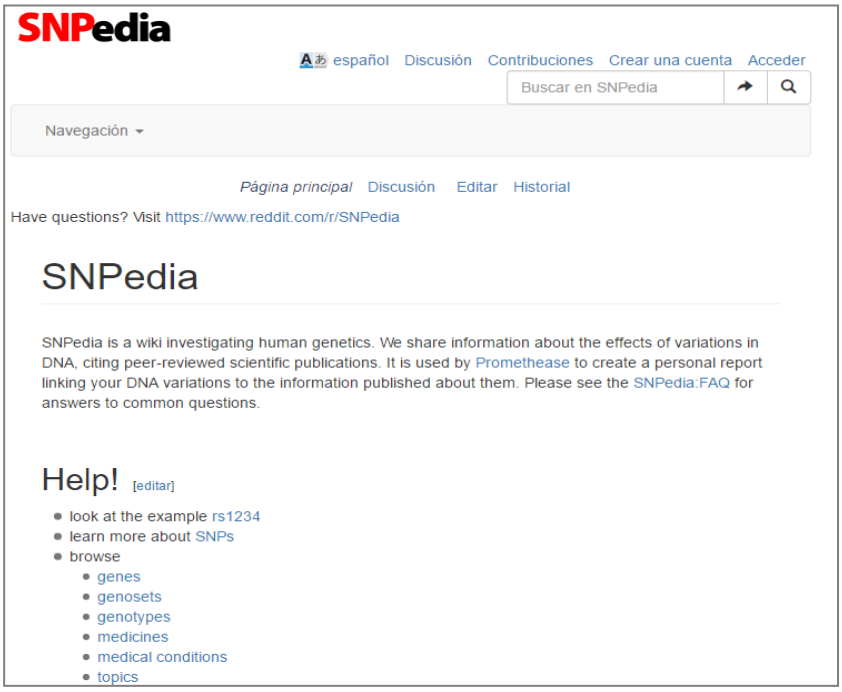


Figura 26. SNPedia (*website*)

Este repositorio está en línea desde 2006, con acceso libre -*disponible*- para uso personal, SNPedia se ha centrado en las asociaciones médicas, fenotípicas y genealógicas de polimorfismos de un solo nucleótido. Las entradas están formateadas para permitir que las asociaciones sean asignadas a genotipos individuales, así como a conjunto de genotipos (*genosets*) [100], [101].

3.2.19 UCSC

El buscador de genomas de UCSC (*University of California Santa Cruz (UCSC) Genome Browser*) ofrece acceso público en línea a una creciente base de datos de secuencias genómicas y anotaciones para una amplia variedad de organismos (Figura 27).

UCSC recopila un conjunto de herramientas integradas para *visualizar*, *comparar*, *analizar* y *compartir* conjuntos de datos genómicos, los cuales se comparten públicamente y son generados por los usuarios [102], [103].

Figura 27. UCSC (*website*)

La base de datos "*UCSC Genome Browser*" posee un gran repositorio de genomas con 166 ensamblajes de *GenBank* [104] que representan más de 93 organismos diferentes a través del árbol de la vida, desde vertebrados como humanos, ratones y pez cebra a insectos y nematodos [105].

3.2.20 UMD (Universal Mutation Databases)

La plataforma UMD (*Universal Mutation Databases*) se desarrolló como un software genérico para crear bases de datos de locus-específicos, con el objetivo de recopilar y analizar los datos generados con el paso de años sobre las mutaciones y su asociación con datos

clínicos y biológicos, los cuales son esenciales para los clínicos, genetistas e investigadores [106].

Figura 28. UMD (*website*)

Esta herramienta está disponible de forma libre a través de su sitio web (*www.umd.be*) (Figura 28). Permite la creación de *LSDBs* para prácticamente cualquier gen e incluye un amplio conjunto de herramientas nuevas para el análisis. Se han implementado nuevas características para integrar: secuencias no codificantes, datos clínicos, imágenes, anticuerpos monoclonales y marcadores polimórficos (SNP) [106], [107].

3.2.21 UniProt (Universalt Protein)

La misión de UniProt (*http://www.uniprot.org/*) es proporcionar a la comunidad científica un recurso completo, de alta calidad y libremente accesible sobre secuencia de proteínas e información funcional (Figura 29) [108].

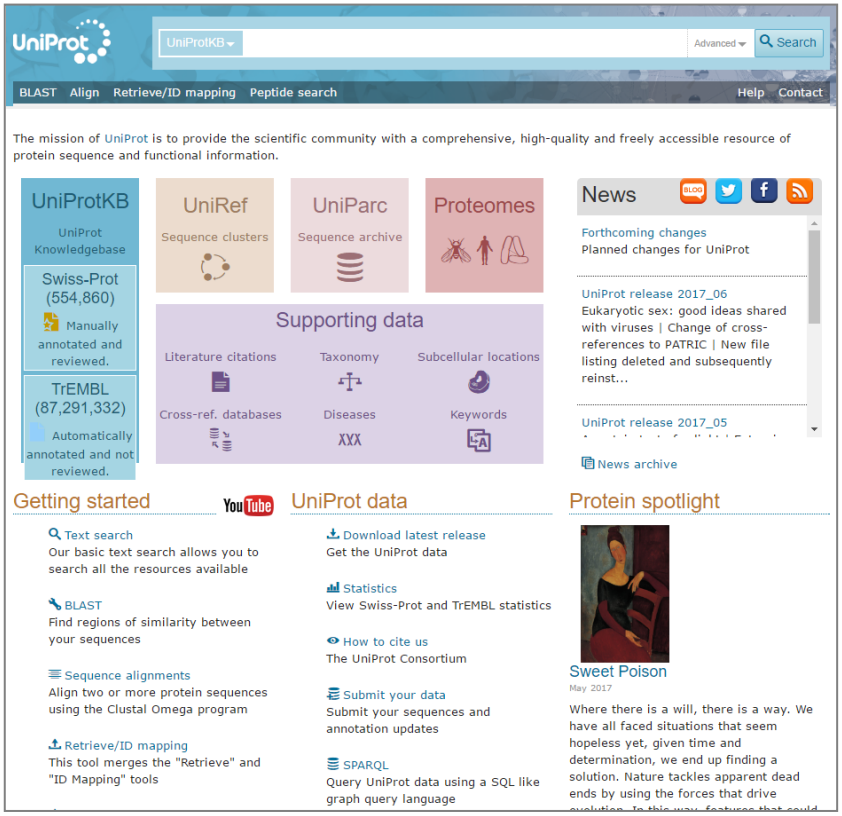


Figura 29. UniProt (*website*)

UniProt se compone de varios elementos importantes.

(1) La sección de UniProt que contiene las entradas chequeadas ("*curadas*") y revisadas manualmente se conoce como *UniProtKB/Swiss-Prot* y actualmente contiene aproximadamente medio millón de secuencias. Estas secuencias crecen a medida que se caracterizan nuevas proteínas.

(2) El resto de secuencias no revisadas se recogen en la sección conocida como *UniProtKB/TrEMBL*. Esta parte de UniProt para el año 2015 contenía alrededor de 80 millones de secuencias y se mantiene creciendo exponencialmente. Aunque estas entradas no se "*curan*" manualmente se complementan con la anotación generada automáticamente.

(3) La base de datos UniParc consiste en un conjunto completo de todas las secuencias conocidas, indexadas por sus sumas de

comprobación[16] de secuencia única y actualmente tiene más de 70 millones de entradas de secuencias.

UniProt tiene referencias cruzadas a más de 150 bases de datos. Actúa pues como un centro de gestión para organizar la información de proteínas [108].

3.2.22 YHRD (Y Chromosome Haplotype Reference Database)

La base de datos YHRD (*Y Chromosome Haplotype Reference Database*) está diseñada para almacenar haplotipos del cromosoma Y, para poblaciones globales. En sus tres versiones anteriores recogía haplotipos para el cromosoma Y, pero en poblaciones especificas (por ejemplo, *europeos*, *asiáticos* y *estadounidenses –de forma separada-*).

El objetivo principal de este proyecto consiste en difundir los datos de frecuencia de haplotipos a: *analistas forenses*, *investigadores* y a todos los que estén interesados en la *genética histórica* y *familiar* [109].

Este repositorio (Figura 30) está estructurado por la asignación de cada muestra de población sometida a un conjunto de poblaciones que comparten un *fondo lingüístico*, *demográfico*, *genético* o *geográfico* común (*meta-poblaciones*). Este principio facilita la evaluación estadística de las coincidencias de los haplotipos debido a un aumento significativo del tamaño de la muestra.

Para la versión 19 (lanzada en agosto de 2006) se podía ejemplificar como en el conjunto de datos de YHRD su rápido crecimiento contribuía a una definición de meta-poblaciones homogéneas (*factible*) sobre las bases de datos genéticas.

Las grandes cantidades de muestra dentro de las meta-poblaciones genéticamente definidas también permiten el desarrollo de métodos bioestadísticos para estimar la frecuencia de haplotipos no observados o raros ("*haplotype frequency surveying method*").

---

[16] *Suma de Comprobación o Checksum*: es un cálculo matemático aplicado a los contenidos de un paquete antes y después de que se envió. Si el "antes" de cálculo no coincide con el "después" de cálculo, hubo errores en la transmisión. Es un dígito que representa la suma de los dígitos en una instancia de datos digitales, que se utiliza para comprobar si se han producido errores en la transmisión o el almacenamiento [198].

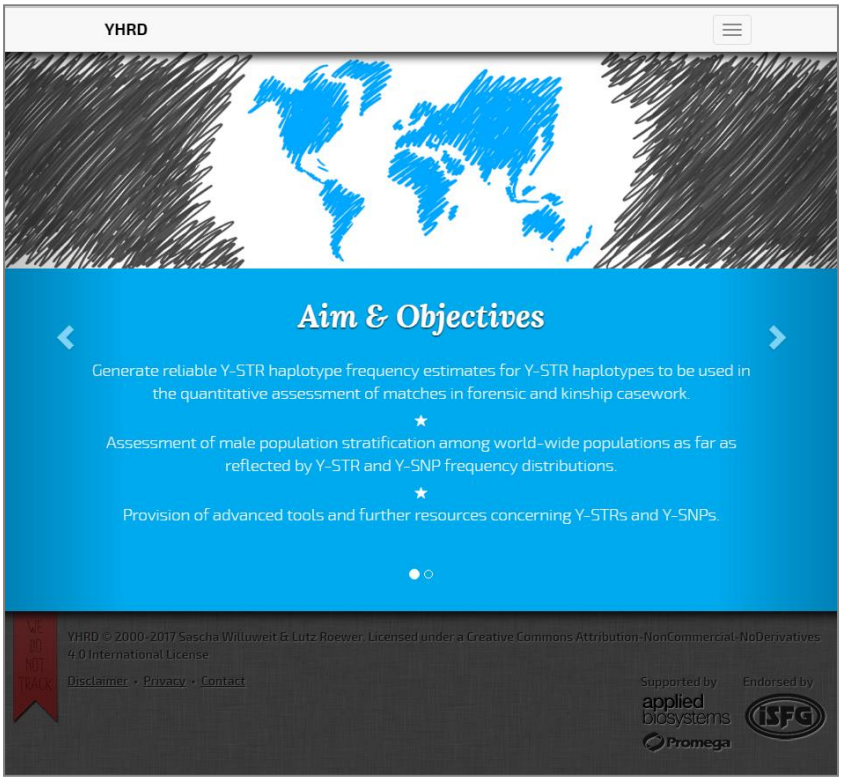


Figura 30. YHRD (*website*)

Para el proyecto YHRD es esencial su carácter colaborativo, en donde impulsan la participación de laboratorios particulares para que hagan sus datos accesibles vía YHRD, permitiendo de esta forma compartir las normas de YHRD con respecto a temas de calidad de los datos [109]. Más información sobre este repositorio se puede consultar en [110], [111].

3.3 Comentarios adicionales

En la Tabla 5 se presenta un resumen de las 22 bases de datos genómicas previamente descritas, indicando a qué se dedica principalmente y algunos detalles específicos.

Hay que destacar dos aspectos esenciales, ligados a la noción de "*caos de datos genómico*" derivada de la información presentada:

1) Con la selección de ese subconjunto de 22 bases de datos se ha querido poner de manifiesto el gran problema de dispersión de datos existente en el dominio genómico, donde existen multitud de fuentes de datos asociadas a distintas granularidades conceptuales asociadas a la información genómica.

Como puede verse en la Tabla 5, a pesar de ser "*solo*" 22, las bases de datos seleccionadas y comentadas –brevemente, pues un tratamiento exhaustivo del contenido y facilidades específicas de cada una de ellas está fuera del ámbito de trabajo de esta Tesis Doctoral- ponen en evidencia esa heterogeneidad en contenidos y formatos de almacenamiento de los datos.

2) No solo es un problema de heterogeneidad: también es un problema de integración y de calidad de los datos manipulados para que su explotación sea correcta. Solo garantizar que bases de datos referidas al mismo tipo de información incluyen los mismos datos es un problema de investigación muy complejo y extenso.

En un entorno como el analizado en el que el ámbito clínico -*medicina de precisión como contexto de aplicación*- es esencial, los errores en la interpretación son simplemente inaceptables, y los mecanismos para asegurar la corrección y la consistencia de los datos son fundamentales.

Tabla 5. Resumen Bases de Datos Genómicas

No.	Base de Datos	Dedicada a…	+Detalles
1	**1000 Genomas**	Catalogar variaciones genéticas	Riesgo de enfermedades del genoma completo (*2,504 personas - 26 poblaciones*)
2	**ALFRED**	Frecuencias alélicas	Orientada a genética de poblaciones y antropología molecular
3	**BIC**	Mutaciones y polimorfismos	Genes de susceptibilidad al cáncer de mama
4	**BioQ**	Consulta/documentación BD genómicas	Herramienta para determinar los orígenes experimentales de los datos
5	**ClinVar**	Variantes genómicas	Interpretar variantes y su relación con la salud humana
6	**COSMIC**	Mutaciones somáticas en el cáncer humano	Datos obtenidos de publicaciones científicas y estudios experimentales a gran escala del Proyecto Genoma del Cáncer
7	**dbGAP**	Genotipo-fenotipo	Gestionar las relaciones e interacciones entre fenotipo y genotipo
8	**dbSNP**	Variaciones genéticas	Posee una colección de polimorfismos genéticos simples
9	**D-HaploDB**	Haplotipos definitivos del genoma	Colección japonesa de CHMs (*complete hydatidiform moles*)
10	**DisGeNET**	Genes y variantes (asociadas a enfermedades)	Datos son extraídos de textos científicos (colección completa)
11	**Ensembl**	Interpretación genómica	Genómica comparativa, regulación transcripcional, etcétera
12	**HapMap**	Haplotipos	Mapa de haplotipos del genoma humano
13	**HGMD**	Mutaciones humanas	Mutaciones de la línea germinal de genes nucleares asociados a enfermedades humanas hereditarias.
14	**KEGG**	Func. Sistema biológico (*célula-organismo-ecosistema*)	Conjuntos de datos moleculares a gran escala
15	**LOVD**	Variación de la secuencia génica/fenotipos humanos	Asociación 3,294 BD de variantes con variantes en pacientes
16	**OMIM**	Genes humanos y desordenes genéticos	Base de conocimiento. Datos extraídos de literatura biomédica
17	**REACTOME**	Traducción de señales, transporte, replicación ADN, metabolismo y otros procesos celulares	Establece relaciones e interacciones entre los procesos biológicos

Tabla 5. Resumen Bases de Datos Genómicas

No.	Base de Datos	Dedicada a…	+Detalles
18	**SNPedia**	Variación genética humana	Asociaciones médicas, fenotípicas y genealógicas de polimorfismos de un solo nucleótido
19	**UCSC**	Secuencias genómicas y anotaciones	Gran repositorio de genomas con 166 ensamblajes de GenBank (*amplia variedad de organismos*)
20	**UMD**	Mutaciones y asociación con datos clínicos/biológicos	Secuencias no codificantes, datos clínicos, imágenes, anticuerpos monoclonales y marcadores polimórficos (SNP)
21	**UniProt**	Secuencia de proteínas e información funcional	Este repositorio actúa como un centro central para organizar la información de proteínas
22	**YHRD**	Haplotipos del cromosoma Y	Su objetivo difundir los datos de frecuencia de haplotipos a: analistas forenses, investigadores y a todos los que estén interesados en la genética histórica y familiar (*para poblaciones globales*)

3.4 Conclusiones

En el presente capítulo se han estudiado los distintos trabajos realizados sobre modelado conceptual y repositorios de datos genómicos, los cuales se relacionan con las contribuciones principales de la tesis doctoral.

Primeramente, se han explicado los principales trabajos desarrollados dentro de la comunidad de modelado conceptual para el entorno bioinformático. En este sentido, es importante destacar que la aplicación de enfoques basados en modelado conceptual permite crear una definición conceptual del dominio de una forma clara y sencilla.

Como se ha observado en los trabajos previos, ninguno de ellos presentaba una solución o una imagen completa del comportamiento del genoma humano, es decir, no contemplaban todos elementos participantes debido a que se enfocaban a procesos específicos del problema global. La solución que se propone en este trabajo define una visión conceptual global, holística de todo el genoma, que se clasifica en distintas vistas del modelo (por ejemplo, *estructural*, *variaciones*, *transcripción*, *rutas metabólicas*, etcétera) con el objetivo de facilitar su comprensibilidad.

Con respecto al tema relacionado con las bases de datos genómicas, se han analizado a modo de ejemplo diversos repositorios (solo un subconjunto considerado relevante dentro de la enorme cantidad de fuentes de datos genómicas existentes) con el objetivo de poner de manifiesto el problema asociado a lo que hemos llamado "*caos de datos genómico*", entendido como la problemática derivada de tener que trabajar en un dominio de conocimiento cuyos datos aparecen dispersos en multitud de fuentes de datos diversas, cada una de ellas especializada en determinada parte específica del conocimiento genómico, pero con carencias muy significativas en todo lo relativo a disponer de una visión conceptual holística, integral, de todos esos distintos ámbitos de conocimiento que conforman dicho dominio genómico.

Hemos evaluado también si los datos almacenados en esas fuentes de datos se apoyan en modelos conceptuales para determinar su forma de representación y su repercusión dentro del entorno bioinformático a través de los datos que gestionan, siendo llamativa la ausencia de modelos conceptuales que permitan realizar una comparación

semántica precisa de sus contenidos. Dentro de los inconvenientes asociados a ese "*caos de datos genómicos*" está la gran heterogeneidad de los datos, provocando redundancia, inconsistencias entre fuentes de datos diferentes, y dispersión de los datos, además de encontrar grandes cantidades de información con mecanismos de gestión con bastante complejidad.

La base de datos (HGDB) presentada en esta Tesis Doctoral busca integrar todo el conocimiento genómico existente actualmente, mediante una estructura basada en modelado conceptual que permita representar y gestionar los datos del dominio de una manera más eficiente.

El objetivo de la aplicación de técnicas de modelado conceptual es generar una gestión de datos más efectiva y eficiente en el dominio genómico debido a i) la mejor *comprensibilidad* de dominio que un modelo conceptual holístico proporciona, ii) la posibilidad de *integrar* con esa perspectiva conceptual unificadora que el modelo conceptual proporciona, y iii) su *versatilidad* para manejar el constante crecimiento del dominio (gracias a las nuevas investigaciones científicas y avances en las tecnologías de secuenciación).

En este ámbito en el que se justifica el principal resultado de esta Tesis Doctoral: un Esquema Conceptual holístico del Genoma Humano como herramienta conceptual esencial de un Sistema de Información Genómica que permita disponer de plataformas que minimicen el problema derivado del caos de datos genómico en el que está instalada la Bioinformática actual.

CAPÍTULO 4

Evolución del Modelo Conceptual del Genoma Humano (MCGH)

¿Por qué es esencial el *modelado conceptual* (MC) [1] para diseñar y desarrollar *sistemas de información* correctos? Esta es una cuestión fundamental dentro de la comunidad de MC, la cual está interesada en demostrar que sólo mediante el uso de técnicas de MC se puede lograr el diseño y desarrollo de *Sistemas de Información* (SI) de calidad.

Para ir un paso más adelante con respecto a esta cuestión, como objetivo principal de esta Tesis Doctoral en la cual se busca responder a esta pregunta de una manera convincente. La necesidad de una estrategia de diseño y desarrollo basado en MC debería ser más evidente mientras mayor sea la complejidad del sistema o dominio en estudio.

La comprensión del genoma humano es un buen ejemplo de un problema extremadamente complejo. El uso del MC para proporcionar una solución que haga frente al caso del "*genoma humano*", se ha explorado inicialmente en trabajos anteriores [45], [46], pero una

perspectiva holística de toda la representación (imagen) todavía no se ha facilitado.

El uso de enfoques avanzados en ingeniería de SI es vital en este ámbito debido a la enorme cantidad de información biológica existente, la cual debe ser *capturada*, *comprendida* (manipulada) y *controlada* de manera eficaz.

Una rama importante de la bioinformática está dedicada a la gestión de los "*datos genómicos*", y la existencia de un gran conjunto de fuentes de datos (diversas) con grandes cantidades de datos que representan un conocimiento relacionado con la evolución continua, lo que hace muy difícil encontrar soluciones convincentes.

Cuando se decide enfrentar este problema desde una perspectiva de SI, es necesario entender que se precisa de la aplicación de MC para comprender la información relevante en el dominio. De esta manera es más sencillo representarla con claridad, lo que permite desarrollar una estrategia de gestión de datos eficaz.

Sin embargo, un punto sorprendente fue el hecho de descubrir que nuestros colegas en biología no tenían idea sobre MC en sus modelos de datos, y la respuesta recibida *–en el mejor de los casos, fue simplemente un diseño lógico relacional-*, una descripción sin una perspectiva de diseño conceptual en lo absoluto.

Por lo que de manera inmediata se llegó a la conclusión de que la perspectiva de modelado conceptual no era para nada utilizada, por lo que se trató de convencer a los colegas del área bioinformática para construir un *Modelo Conceptual del Genoma Humano* (MCGH) con el objetivo principal de *-comprender como entendemos que funciona nuestra vida en la tierra-*, permitiendo proporcionar una comprensión de los conceptos básicos que explican como la estructura genotípica se manifiesta en un fenotipo externo.

El objetivo se basa en demostrar la necesidad de un MC:

- Compartir la comprensión de los conceptos esenciales del dominio *-en nuestro caso el genoma humano-*, y
- Orientar el diseño y el desarrollo de las correspondientes bases de datos (BD), que normalmente cubren sólo una parte de los MC. Esto significa que el uso del modelado conceptual sólo como un tipo holístico *-bases de datos conceptual-*, hará posible

la integración de diferentes fuentes de datos que representan diferentes perspectivas del conocimiento genómico.

Para presentar todo el trabajo realizado siguiendo esta línea de investigación, en primer lugar, se introduce una primera representación conceptual del conocimiento genómico correspondiente, el cual denominamos Modelo Conceptual del Genoma Humano versión 1 (MCGH v1) (Sección 4.1). Después de este ejercicio conceptual, se generaron más debates en profundidad sobre cómo representar mejor los conceptos básicos de este dominio, en donde el conocimiento asociado se mantiene cada día en constante evolución.

Como resultado de este trabajo conceptual se presentó una extensión a la versión inicial, identificada como ECGH v1.1 (Sección 4.2) hasta llegar a la última versión propuesta del modelo conceptual, llamada MCGH v2 (Sección 4.3 y 4.4). Finalmente, la Sección 4.5 presenta las conclusiones del capítulo. Los resultados de este capítulo se encuentran publicados en los siguientes trabajos [112] y [7]

4.1 Modelo Conceptual del Genoma Humano, versión 1 (MCGH v1)

La primera versión del esquema se caracteriza por ser el primer intento en abordar la descripción holística del dominio de la genómica, y como tal se centra en una visión del genoma centrada en sus conceptos más básicos, obviando algunos aspectos más complejos.

El MCGH versión 1 centra su atención en el análisis de genes individuales, sus mutaciones y sus aspectos fenotípicos. En consecuencia, otros fenómenos como:

- *la regulación múltiple,*
- *la codificación de una misma proteína por dos genes diferentes,*
- *los pseudo-genes, o*
- *la acción combinada de múltiples genes*

Son apartados con la intención de ser estudiadas en futuras versiones del modelo. Esta primera versión se podría considerar como la "*esencial*".

Para concretizar la primera versión del MCGH se reunió a un amplio grupo de expertos en las áreas de *Biología Molecular* y *Modelado Conceptual*, con el fin de abordar los elementos principales y esenciales

del dominio genómico. Teniendo en cuenta de que a medida en que se extendiera el conocimiento (nuevos) sobre el domino se podría generar "N" versiones del MCGH.

En las etapas de modelado se desarrollaron cuatro iteraciones para llegar a la versión 1 del MCGH (*las cuales se pueden consultar en el siguiente trabajo* [113]). A continuación, se presenta la clasificación de la primera versión en tres vistas principales:

- *Gene-Mutation View*: Utilizada para modelar el conocimiento sobre los genes, su estructura y sus variantes alélicas. Las entidades con mayor relevancia que han sido modeladas en esta vista son: "*Gene*" y "*Allele*" (Figura 31).
- *Genome View*: Esta vista se encarga de modelar genomas humanos individuales. Es una vista vital para la futura extensión del modelo (Figura 32).
- *Transcription View*: Esta vista modela los componentes básicos del proceso de transcripción y las síntesis de las proteínas (que es lo que se conoce como "*expresión génica*") (Figura 33).

4.1.1 Gene-Mutation View

Para representar el concepto esencial de "*Gen*", se supone que las diferentes variantes posibles –denominadas "*alelos*"- para un gen, están asociadas a una clase "*Gene*" primaria mediante el uso de una clase "*Allele*" que posee dos especializaciones:

1. Una que representa la secuencia considerada de referencia para el gen, y
2. Otra representa una posible versión variada con respecto a la anterior.

La secuencia de referencia contiene un atributo "*sequence*" que obtiene la secuencia de ADN especifica de un gen. Las versiones variadas del gen que se incluyen en la clase "*Allelic Variant*" utilizan un atributo derivado que representa un cambio en la secuencia de referencia.

¿Qué estructura tiene un alelo? ¿Es una estructura suficientemente bien definida para ser incluida en nuestro modelo conceptual? Se debe explorar esta interrogante con un mayor nivel de detalle para poder almacenar información genética de "*valor*". El primer proceso –*esencial*- a nivel biológico y que caracteriza al comportamiento génico, es el proceso de transcripción, y la respuesta que se facilita a esta cuestión se representa mediante el concepto de "*Transcription Unit*".

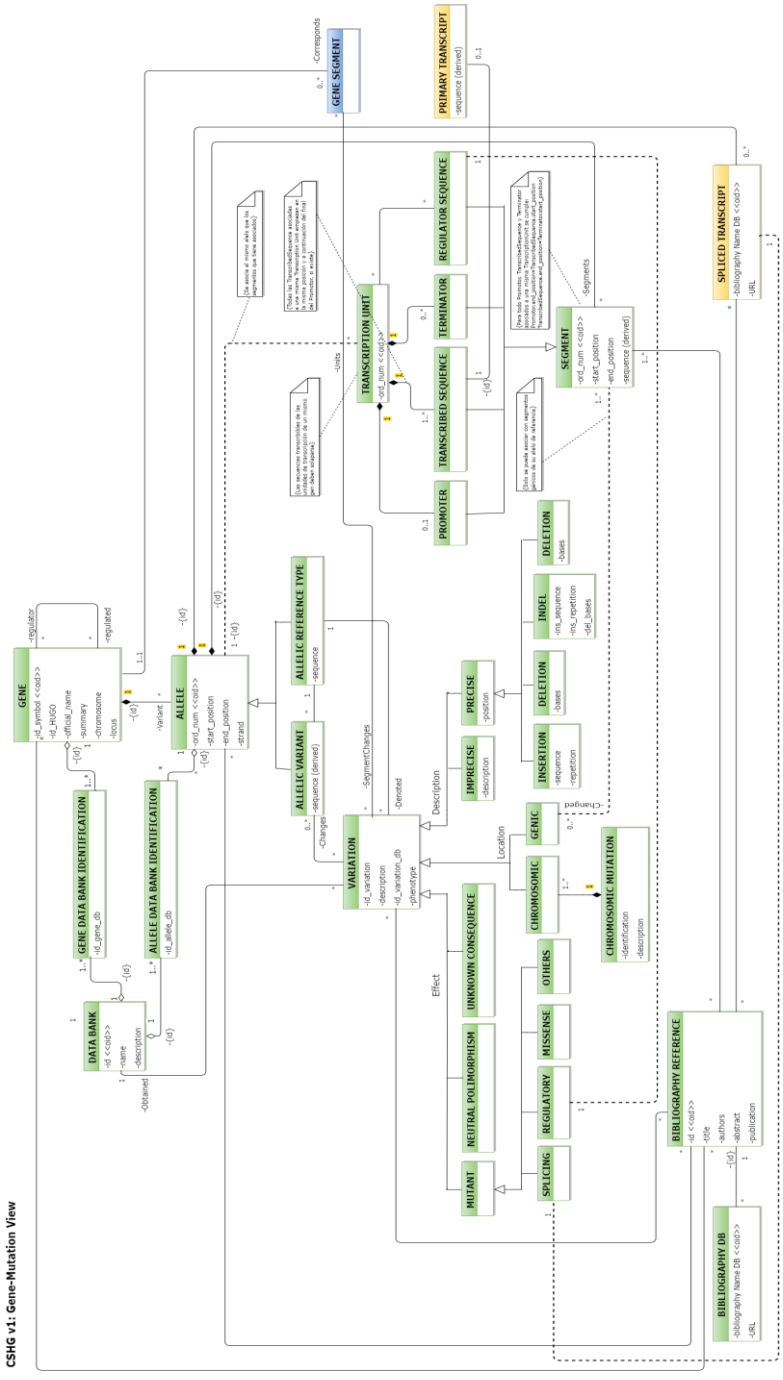


Figura 31. MCGH v1: "*Gene-Mutation View*"

Esto se representa a través de una asociación entre la versión alélica seleccionada de un gen, y el conjunto de piezas de ADN que lo componen –llamado "*Segments*" en el modelo-. La estructura de un "*Segmento*" contiene los siguientes componentes:

- *Un promotor*: el cual describe la secuencia de ADN, marca el comienzo de la transcripción.
- *Secuencias transcriptibles*: es responsable de describir la secuencia de ADN transcrita por la ARN polimerasa.
- *Un terminador*: describe el final del proceso de transcripción en la secuencia de ADN.
- *Secuencia reguladora*: describe un segmento alélico que contiene las secuencias de nucleótidos de las funciones reguladoras de uno o más procesos de transcripción.

Para caracterizar en detalle las posibles variantes, el modelo conceptual introduce una clase "*Variation*", en donde se especifican los cambios aplicados o detectados sobre la secuencia de referencia. El siguiente punto importante consiste en identificar los tipos de variaciones que se consideran relevantes. Concretizando lo que se conoce actualmente, se clasificaron en tres tipos principales de variaciones:

- La primera, centrada en el *efecto* que tiene una variación (es una mutación asociada a una enfermedad) o simplemente es una variación neutra *–en el sentido de que no tiene un efecto negativo en términos clínicos-* (en el lenguaje biológico, se refiere como "*Polimorfismo Neutro*").
- El segundo considera la *ubicación* de la variación, por ejemplo, si es una variación cuyo alcance está en el cromosoma o en el gen.
- La tercera considera si la variación tiene una *descripción* asociada con ella. La descripción podría tomar dos formas: (a) una variación "*precisa*", cuando la estructura y los nucleótidos involucrados son claramente conocidos, o (b) si la estructura de una variación no se conoce, entonces se habla de una variación "*imprecisa*", cuyos efectos están probablemente aún por descubrir.

Es importante fijar cuantos tipos de cambios pueden presentar las variaciones precisas, en esta primera versión del MC se manejan las siguientes [114]:

- *Inserciones*: ocurren cuando las variantes alélicas presentan inserciones de nucleótidos con respecto al alelo de referencia.

- *Deleciones*: cuando en el detalle de las variantes alélicas detectamos que ciertos nucleótidos han sido borrados en una posición específica del alelo de referencia.
- *Indel*: este caso se encuentra cuando en las variantes alélicas ocurre un borrado de nucleótidos y una inserción de uno o varios nucleótidos, un determinado número de veces en la misma posición de la cadena.
- *Inversión*: describe los detalles de las variantes alélicas en las que ocurre una inversión en el orden de los nucleótidos con respecto a la secuencia de referencia.

En esta versión del modelo conceptual, el concepto de *SNP*[17] no fue representado explícitamente porque se consideró como una variación normal. En el desarrollo del MCGH v1 se detectó un punto importante a considerar: el uso de secuencias de referencias de fuentes externas. Desafortunadamente, los identificadores de genes a veces dependen de la fuente de datos externa. Por lo tanto, es importante representar en el modelo conceptual las fuentes de datos externas que se han utilizado *–para la identificación de un gen y un alelo-* (representado en el MC, por medio de las clases "*Gene Data Bank Identification*" y "*Allele Data Bank Identification*" respectivamente).

También es importante destacar que esta vista del modelo representa el origen de los datos considerados como "*relevantes*", por sus implicaciones en el campo de la salud. El objetivo es apoyar el conocimiento con datos bibliográficos, y presentar así una visión más clara de las fuentes de conocimiento y sus datos asociados, como, por ejemplo, títulos de artículos, autores, publicaciones asociadas, nombre del repositorio y cualquier otro dato de interés, siempre vinculando esta información al *gen*, *alelo* o *variación* tratada.

4.1.2 Genome View

La siguiente vista (Figura 32) se introduce con el objetivo de incluir genomas individuales, los cuales podrían ser comparados estructuralmente con la vista anterior *Gene-Mutation*.

[17] Un polimorfismo de un solo nucleótido (SNP) es una variación en una única posición en una secuencia de ADN entre los individuos. Recordemos que la secuencia de ADN se forma a partir de una cadena de cuatro bases de nucleótidos: A, C, G y T [*Scitable by Nature Education*].

Esta vista proporciona una perspectiva general relacionada con toda la noción y caracterización de la composición concreta del genoma. De hecho, el genoma está compuesto de cromosomas –en el caso de los humanos "*23 pares*"-. Estos están representados en el MC mediante la clase "*Chromosome*".

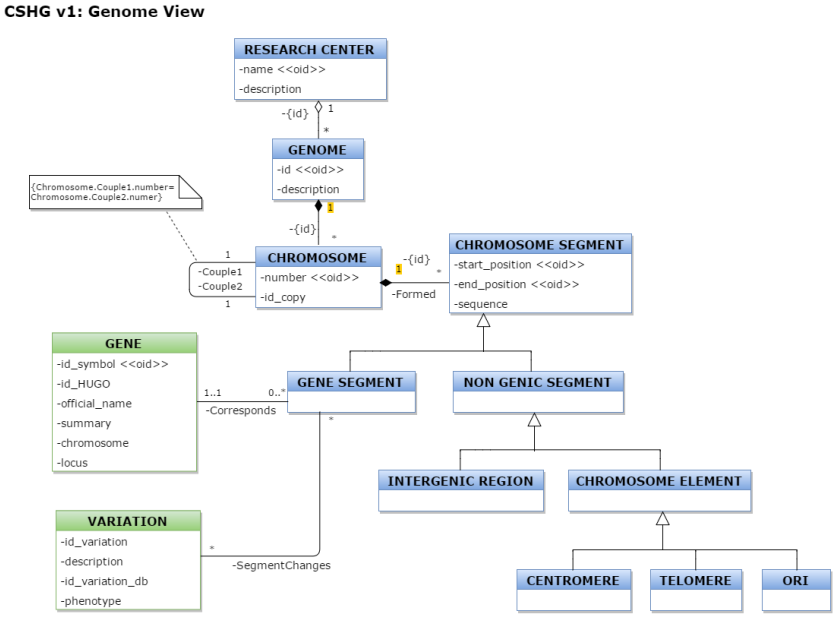


Figura 32. MCGH v1: "*Genome View*"

Las secuencias de los cromosomas son largas, y para lograr manejarlas con un criterio de unidad funcional, se dividen en "*piezas*" más pequeñas (componentes con una identidad funcional). En este trabajo se utiliza la noción de "*Segment*" para explicar este caso. La introducción de este elemento como un componente básico de una secuencia genómica global nos permite identificar qué partes específicas del ADN cromosómico tienen un significado importante. La composición de todos estos segmentos cromosómicos representa la secuencia de todo el cromosoma.

Los segmentos de los cromosomas pueden ser de dos tipos: "*Coding*" y "*Non-coding*", dependiendo de si están o no asociados a la síntesis de proteínas. En el modelo conceptual se encuentran etiquetados como "*GenicSegment*" y "*NonGenicSegment*". Los segmentos génicos representan las partes codificantes que tradicionalmente se consideran

más significativas para el cromosoma y que están relacionadas con los genes. Pero cada vez está más claro que los componentes de ADN "no génicos" también tienen funciones vitales para explicar la operación genómica. Esta es la razón por la cual existe la necesidad de distinguir entre los segmentos génicos (relacionados con un gen y conectado a través del concepto de "*gen*" con la vista anterior "*Gene-Mutation*") y los segmentos no génicos.

Del mismo modo, es necesario distinguir entre los diferentes tipos de segmentos no génicos. Conforme al conocimiento actual, se identificaron dos tipos: (a) las regiones intergénicas que representan el espacio entre los genes, y (b) las que forman parte de la estructura de los elementos cromosómicos. Entre estos elementos cromosómicos, detectamos tres elementos cromosómicos de interés:

- *Centrómero*: es la región condensada o constreñida que separa el brazo corto del brazo largo en el cromosoma.
- *Telómero*: se le denomina al extremo de cada brazo.
- *ORI*: representa el origen de replicación.

Una de las ventajas derivadas del uso de un modelo conceptual es su gran facilidad de adaptación para nuevos conceptos provenientes de la evolución del conocimiento en el dominio. Esto significa que, si se identifican nuevos elementos en el dominio genómico, incluirlos como nuevos componentes del MC en el lugar correcto debería ser siempre un ejercicio no complejo (complicado).

Esta vista hace posible la representación de genomas completos (individuos) y permite la introducción de información asociada a dos elementos importantes: (a) los centros o institutos de investigación responsables de secuenciar los genomas, y (b) el resultado final del proceso de secuenciación de una muestra dada (representado en el modelo conceptual mediante las clases "*Research Center*" y "*Genome*" respectivamente).

4.1.3 Transcription View

Tras finalizar la explicación de la representación conceptual de la estructura general de "*genes*" y la colección de "*genomas*" (individuales), se procedió a especificar las partes que conforman la vista de transcripción dentro del modelo conceptual.

El objetivo principal de esta vista es representar el proceso de "*síntesis de proteínas*", mediante la integración de una definición de la estructura interna de los alelos, con el fin de describir cómo los elementos antes mencionados están involucrados en el proceso de transcripción de ADN. El primer aspecto estudiado fue la representación del ARN transcrito de la copia de ADN de la secuencia transcribible (relacionada con una "*Transcribable Sequence*" y, por lo tanto, vinculada a la vista "*Gene-Mutation*").

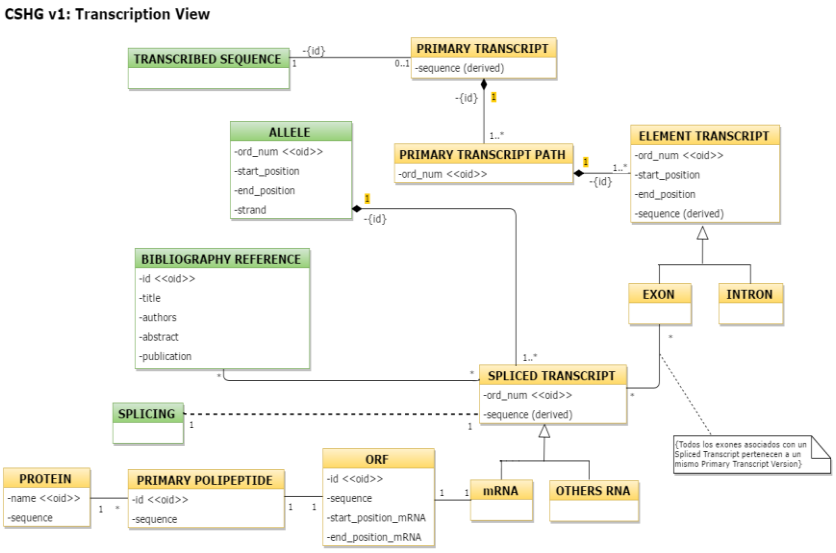


Figura 33. MCGH v1: "*Transcription View*"

Este producto de ARN que se obtiene inmediatamente después de la transcripción se conoce cómo "*Transcripción Primaria*" en el vocabulario biológico. La transcripción primaria está constituida por una o más particiones (representadas en el MC como componentes principales de una clase llamada "*Primary Transcript Path*"). Cada partición tiene dos tipos de elementos de transcripción: "*exones*" e "*intrones*".

Los exones presentan combinaciones diferentes para una cierta partición de la transcripción primaria. En el MC se representan las diferentes combinaciones de exones mediante la clase "*Spliced Transcript*" (la cual está relacionada con un "*Allele*", y por lo tanto se relaciona nuevamente con la vista "*Gene-Mutation*"). El proceso de "*Splicing*" se basa en la eliminación de intrones y la unión de exones en

el mRNA antes de salir del núcleo. Los resultados de este proceso son representados mediante la clase "*Splicing Transcript*", pudiendo producir dos resultados de *Splicing* diferentes: (a) el "*mRNA*", y (b) el "*Alternative Splicing*" (representados en el modelo conceptual a través de la clase "*Others RNA*").

El "*mRNA*" (*ARN mensajero –ARNm-*) es el resultado de la transcripción de un gen y lleva la información necesaria para sintetizar una proteína. Para completar la vista de transcripción, se debe modelar el camino desde el ARNm hasta el proceso de traducción de proteínas resultante. Dentro del ARNm encontramos los "ORF" (de sus siglas en inglés, "*Open Reading Frame*" [115]). Después de terminar la traducción de un ORF, se genera una cadena de aminoácidos por la estructura primaria de la proteína (clase "*Primary Polipeptide*"). Las transformaciones químicas de la cadena de aminoácidos producen como resultado final una proteína funcional (representada en el modelo conceptual mediante la clase "*Protein*").

La combinación de estas tres vistas conforma la primera versión del Modelo Conceptual del Genoma Humano (MCGH v1).

4.2 MCGH versión 1.1

El Modelo Conceptual del Genoma Humano versión 1.1 (MCGH v1.1), es la evolución natural del MCGH v1. Este básicamente comprende la inclusión de la vista *fenotípica* en el modelo. La visión "*fenotípica*" es muy importante porque aporta mayor consistencia al modelo.

El hecho de ofrecer una visión "*genotípica*" (información genética que posee un organismo), ligada a una visión "*fenotípica*" (expresión del genotipo en función de un determinado ambiente), ofrece un gran valor investigativo y dota de mayor importancia al modelo.

Algunas nociones sobre los conceptos que fueron descritos y modelados en la versión 1 han sufrido modificaciones debido a la continua evolución del dominio. Un ejemplo son los nuevos conocimientos descubiertos y/o validados.

4.2.1 Phenotype View

¿Cuál es la contribución más importante realizada en esta extensión del modelo? El punto más importante para esta versión es la inclusión de una nueva "*vista*" relacionada con el "*Fenotipo*".

La extensión realizada en el modelo conceptual con la introducción de la "*Vista Fenotípica*" otorga una gran relevancia al modelo al momento de representar variaciones con el fenotipo al que están relacionadas. Hoy en día, la visión conjunta de "*genotipo-fenotipo*" es una de las áreas más investigadas y con gran importancia en el dominio genómico.

¿Qué es el fenotipo? Se conoce como "*Fenotipo*" a cualquier característica o rasgo observable de un organismo [116]. El fenotipo es la expresión del genotipo, o conjunto de genes de un individuo, de acuerdo con un entorno dado y la interacción entre ambos (Figura 34).

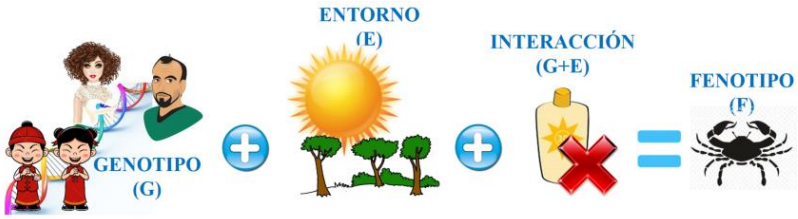


Figura 34. Genotipo y Fenotipo

Se llama "*genotipo*" de un organismo al conjunto de instrucciones heredadas, que el individuo lleva en su código genético. Sin embargo, no todos los organismos con el mismo genotipo se aprecian o actúan de la misma manera, en la apariencia y en el comportamiento de estos debido a las influencias del entorno (ambiental) y de desarrollo.

El fenotipo se compone de todas las características a simple vista, las visibles mediante procedimientos técnicos y por el comportamiento del individuo (por ejemplo: color rubio, tipo sanguíneo "A+" o un comportamiento compulsivo).

Hay un término médico llamado "*Syndrome*", el cual se compone de un conjunto de características fenotípicas (Figura 35) [117]. El concepto de "*síndrome*" se define según la RAE, como "*un conjunto de síntomas característicos de una enfermedad o un estado determinado*" (este

concepto es representado en el modelo mediante la clase "*Syndrome*" para todo lo que posea un carácter patológico).

Un síndrome puede tener características severas (patológicas) asociadas, de modo que las "*Variaciones*" clasificadas con mutaciones estarían relacionadas con la clase "*Syndrome*" (la vista fenotípica se conecta con la vista "*Gene-Mutation*" a través de la clase "*Variation*").

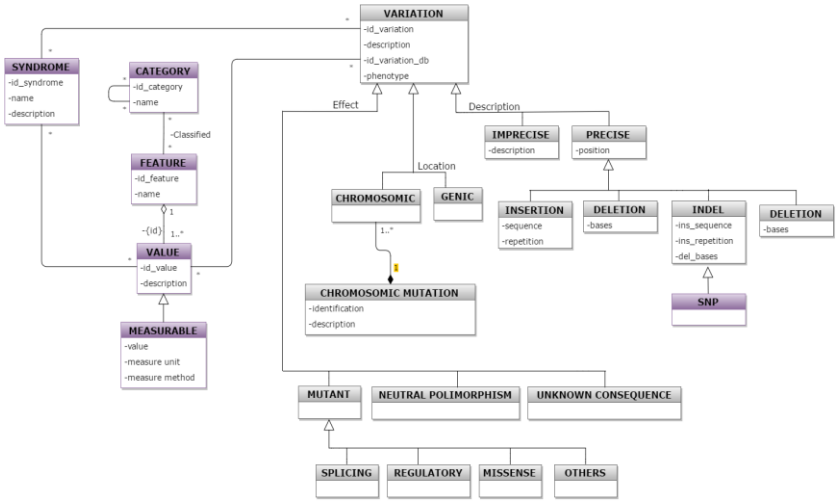


Figura 35. MCGH v1.1: "*Phenotype View*"

Actualmente existen enfermedades asociadas a múltiples características, y para modelar estas características en el MC, se han clasificado sus elementos en diferentes niveles que se definen a continuación:

- *Característica*: es una cualidad particular que permite distinguir algo con otros de la misma especie. Para ilustrar este caso podemos tomar como síndrome, el caso de la –*neurofibromatosis*-. Este síndrome tiende a producir manchas en la piel, y con el objetivo de representar estas características en el MC se incluye la clase "*Feature*".
- *Valor*: este concepto representa una o varias cualidades (positivas o negativas) de la característica descrita. Es importante representar el valor de dicha característica. Las características pueden tomar diferentes colores o tamaños como un valor, y éstos se representan en el MC mediante la clase "*Value*".

- *Categoría*: este concepto se define para clasificar las características de acuerdo con el órgano en que se encuentra o afecta. Las categorías pueden contener varias subcategorías. Estos conceptos se representan en el MC a través de la clase "*Category*", la cual está asociada a la clase "*Feature*" (las subcategorías están representadas en el MC por una relación reflexiva en la clase "*Category*").
- *Medida*: este término se introduce debido a la importancia de modelar la manera en que las características se miden o perciben para dar un diagnóstico. Por ejemplo, si se quiere decir que el tamaño de un –neurofibroma- es "27", por si solo es un dato ambiguo (no es lo mismo que decir "27mm" y "27cm"). En el MC los valores medibles son representados por una extensión de la clase "*Value*", a través de su clase hija "*Measurable*".

Un aspecto importante para considerar es que no todas las características fenotípicas tienen que estar relacionadas con un síndrome (es decir, no tienen que ser patógenas). El color de los ojos verdes, por ejemplo, puede ser una característica particular y no tiene relación patógena.

En el modelo conceptual con el objetivo de abordar todos los posibles casos, se ha relacionado la clase "*Value*" directamente con la clase "*Variation*", permitiendo que un valor esté asociado con muchas variaciones y una variación con muchos valores, como se muestra en esta versión del modelo conceptual (Figura 35).

Otra contribución importante de esta extensión es la inclusión en el MC del concepto de "*SNP*", del acrónimo en inglés "*Single Nucleotide Polymorphism*". Un SNP se define como una *variación en la secuencia de ADN que implica una única base de la secuencia del genoma*. Además, para tal variación será considerada como SNP si posee una frecuencia de al menos 1% en una población dada, sino se clasifica como una variación puntual [118].

La incorporación del concepto de SNP en el modelo conceptual es esencial porque a través del mismo se obtienen los medios para diferenciar distintos tipos de polimorfismos. La diferencia más notable es la separación resultante entre las variaciones genotípicas patógenas o no (esto se debe a que las variaciones del SNP se consideran "*polimorfismo neutro*").

El "*SNP*" está representado en el MC como un tipo de especificación "*Indel*", siendo así un concepto heredado de la clase "*Variation*". Por otra parte, es importante resaltar el hecho de que un SNP se puede detectar en diferentes poblaciones (con niveles de ocurrencia porcentualmente distintos). Esto da lugar a que otros polimorfismos, mutaciones o variaciones con efecto desconocido, posean diferentes efectos con respecto a la población a la que pertenecen. Este término se representa en el MC mediante la clase "*Population*", la cual está relacionada con las clases "*Variation*" y "*SNP*".

4.3 Desde v1 a v2: MCGH v2

Esta sección explica los pasos realizados para pasar de la versión 1 del MCGH hasta la versión 2. El Modelo Conceptual del Genoma Humano versión 2 (MCGH v2) cambia su núcleo central y pasa de representar una visión "*gencentrísta*" a una visión centrada en el concepto de "*cromosoma*". Este cambio de visión en el modelo representa la principal diferencia con respecto a las versiones anteriores del modelo (v1 y v1.1).

Una vez que lo que se denominó MCGH v1 se consideró finalizado, se empezó a evaluar su capacidad para hacer frente a los datos reales que se manipulan en el dominio bioinformático. En el momento en que se decide poner en práctica la versión inicial del MCGH, identificamos un conjunto de preguntas para abordar:

1. No se estaba seguro sobre la conveniencia de combinar una vista genómica relacionada con el almacenamiento de genomas individuales -la llamada "*Genome View*" en v1-, con una vista genómica estructural más teórica relacionada con la configuración y caracterización del genoma en su conjunto –la llamada "*Gene-Mutation*" y "*Transcription*" *View*-.
2. Con respeto al concepto central del gen, no siempre es factible describir la estructura del ADN en términos de genes como constructores básicos. En este punto, se llegó a la conclusión de que la estructura más adecuada sería la utilización de los "*elementos cromosómicos*" como los bloques de construcción básicos. Cuando el trabajo se centró en los -elementos cromosómicos- como secuencias de ADN especificas con un comportamiento relevante identificado para la maquinaria

celular, se debería ser capaz de describir la estructura del genoma con más detalle.

3. La incorporación de información relevante y nueva con mayor nivel de detalle en el MCGH es una necesidad esencial, especialmente cuando los conceptos básicos están involucrados en la discusión. Es importante integrar conceptos más relevantes, como, por ejemplo, el concepto de SNPs. Sobre el cual ya existe información que debe ser incluida apropiadamente en el modelo conceptual. En esta sección se explica cómo abordar el problema de incorporar este conocimiento en la versión actual del MCGH, y como el análisis conceptual del problema abre la posibilidad de introducir nuevas mejoras para modelarlo mejor.
4. Se detectó la necesidad de extender la versión 1 con información más significativa asociada con el genoma. Concretamente, pasar de genotipo a fenotipo de una manera completa y unísona, pues al desarrollar la versión 1 se omitió la especificación (desde la perspectiva y descripción) de los "*Pathways*".

El desarrollo de estas cuatro ideas llevó a la evolución del Modelo Conceptual del Genoma Humano versión 2, la cual será explicada en detalle a continuación.

4.3.1 Eliminación del banco de datos de genomas individuales

Al revisar el conocimiento representado en el MCGH v1, el modelo (*plantilla*) genérico del genoma *–el cual es la estructura precisa del genoma humano y su caracterización-* y la perspectiva de los bancos de datos del genoma *–sobre como almacenar los genomas individuales objetos de estudio-* se realizó de una forma mixta: las vistas de *Gene-Mutation* y *Transcription* aparecen juntas en el modelo conceptual. Para representar con mayor precisión el conocimiento del dominio, las propiedades genéricas y muestras individuales del genoma deben distinguirse claramente. Si se separa la muestra (individual) de un paciente de la base del genoma (plantilla) tomada como referencia, sería más fácil encontrar, por ejemplo, variaciones significativas asociadas a enfermedades.

El MCGH v2 omite por lo tanto la denominada "*Genome view*", centrándose en una descripción más precisa sobre el modelo genérico

del genoma para recopilar toda la información relevante sobre el genoma. En esta versión se decidió organizarlo en cinco vistas principales [112]:

1. *Structural view*: esta vista se compone de los elementos básicos de la secuencia de ADN (describe la estructura del genoma),
2. *Transcription view*: esta vista posee los componentes implicados en el paso del ADN a la diversidad de ARNs (muestra los componentes y conceptos relacionados con la síntesis de proteínas),
3. *Variation view*: esta vista se encarga de caracterizar los cambios en la secuencia de referencia, las cuales tienen implicaciones funcionales en la forma en que se expresa el genoma (describe los cambios en la secuencia de referencia),
4. *Pathways view*: esta vista pretende enriquecer el modelo conceptual con información sobre las rutas metabólicas para unir los componentes del genoma que participan en las rutas con expresiones fenotípicas (describe información sobre las rutas metabólicas),
5. *Bibliography and data bank view*: esta vista se encarga de evaluar la fuente de cualquier información con el objetivo de establecer de donde proviene cualquier dato.

4.3.2 Los elementos cromosómicos como unidades básicas de modelado

El uso de elementos cromosómicos como elementos básicos para la construcción del ADN tiene una influencia directa en la forma en que las variaciones y su origen de ADN se representaron en el modelo conceptual (MC). En la versión 1, la noción de "*alelo*" fue representada como una noción derivada explícitamente –a través de la clase "*Allelic Variant*"-. Además, las variaciones fueron relacionadas con segmentos génicos, ya que no fue posible registrar variaciones cuya fuente se encontraba en otras partes del genoma –*no génico*-. Para superar este problema, la propuesta conceptual consistía en relacionar directamente una variación con una posición cromosómica (especifica) del ADN, debido a que esta solución representa mejor la estructura real del genoma.

El beneficio de no tener la variación directamente relacionada con una variante alélica es doble:

1. En primer lugar, permite definir la variación con mayor precisión, ya que está asociada con una secuencia única del genoma donde se produce la variación. De esta manera, la variación no es dependiente de la variante alélica y la correspondiente asociación "*mucho-a-muchos*" como se planteó en la versión anterior. Realmente esta relación era un problema y conducía a la siguiente cuestión: *¿Cómo determinamos por ejemplo que las diferentes variaciones de un alelo común no eran incompatibles*?
2. En segundo lugar, el concepto de "*variante alélica*" ya no se necesita explícitamente. Como no se tienen genomas individuales en el modelo, se elimina la necesidad de manejar variantes alélicas. A medida que el conocimiento sobre el dominio genómico fue mejorando y aumentando, surgió la pregunta de que si las variantes alélicas de referencia realmente existen. Porque de ser así, significaría que hay un catálogo de variantes bien determinadas cuya estructura y comportamiento deben ser conocidos perfectamente.

La introducción de este conocimiento en el modelo podría llevarse a cabo en cualquier momento. Pero, aunque no existe una respuesta precisa para esta cuestión, se puede concluir con que la omisión de la clase "*Alelic Variant*" permite obtener una descripción más clara, conceptualmente hablando.

Ahora bien, en cualquier caso, es posible generar instancias de alelos utilizando las combinaciones adecuadas de variaciones, porque puede ser vista como información derivada, obtenida mediante la aplicación de un conjunto de variaciones seleccionadas en la secuencia de referencia. Presentar instancias de una clase "*Alelic Variant*" implica caracterizar el conjunto especifico de variaciones que "crean" el alelo considerado.

En este trabajo se argumenta que esta representación (versión 2) es más precisa porque la clasificación de estas preocupaciones conceptuales se hace explicita, el modelo conceptual está en un estado *–semánticamente hablando-* claro, y permite incorporar nuevos conocimientos. Mediante esta solución se proporcionan respuestas satisfactorias a las preguntas abiertas que están provistas en el proceso de comprensión del genoma. Siempre que se identifiquen un conjunto de variaciones como un todo semántico, un alelo sería el resultado de aplicar este conjunto especifico de variaciones a la secuencia de

referencia que forma un elemento cromosómico como una cadena de ADN de nucleótidos. La representación de este conocimiento "*alélico*" se deja de lado hasta una siguiente versión del modelo conceptual.

4.3.3 Modelado de SNPs

En la versión inicial, un concepto genómico de alta relevancia como es el "*SNP*" (*Single Nucleotide Polymorphism*) no se encontraba explícitamente representado en la definición conceptual. La especialización de las diferentes variaciones realizadas en la versión 2 es mucho más precisa, ya que distingue entre dos categorías: (1) la *frecuencia* de la variación, y (2) su *descripción* –precisa o imprecisa-.

Yendo más allá de esta simplificación conceptual, es importante tener en cuenta como se almacenan los SNPs actualmente en las distintas fuentes de datos, como, por ejemplo en *dbSNP* (utilizada ampliamente) [71]. Consultando sus representaciones actuales, se realizó un ejercicio de ingeniería conceptual inversa para incluir en el MCGH un conjunto de clases que representen y definan este conocimiento.

Para ello, se describió que un SNP se especifica en este dominio como un conjunto potencial de variaciones en las que un nucleótido cambia por otro. Este cambio permanece abierto, lo que significa que la noción de variación en este caso se puede definir como una posición en la secuencia de referencia que puede tener valores diferentes dependiendo de la población estudiada, los cuales incluirían una frecuencia asociada (dada). Este caso ha sido abordado usando una jerarquía de especialización para los SNPs. Por lo que este cambio generó una discusión. Cualquier variación precisa se modela como una variación individual donde la secuencia de referencia "*sufre*" un cambio.

Sin embargo, la forma en que los SNPs son tratados es algo diferente: un SNP define que nucleótido se altera. Aparece en la secuencia de referencia de origen (a través del atributo "*alelo*" para el caso homocigótico, "*alelo1*" y "*alelo2*" para el caso heterocigótico). Esta representación conserva la forma en que los datos relacionados con el SNP aparecen en la configuración genómica real. Pero la visión de los SNPs como un conjunto de variaciones individuales sugiere que una mejor representación podría ser modelar el SNP como una agregación de variaciones precisas (*indel*).

Este es un tema para una discusión más abierta, y se podría considerar para una próxima evolución del MCGH. Este cambio podría representar mejor –*conceptualmente*- lo que es un SNP, pero el cambio tiene que ser cuidadosamente analizado porque la gestión de datos de los repositorios de datos actuales de SNPs debe ser adecuadamente adaptada a la nueva representación de datos.

4.3.4 Introducción de los conocimientos relacionados con: Pathways

Una de las innovaciones importantes de esta versión es la extensión del modelo conceptual con la integración de los "*Pathways*". Dentro de los *Pathways* biológicos más importantes se encuentran tres tipos principales:

1. *Metabólicos*: estas rutas hacen posibles las reacciones químicas que se producen en el organismo, como, por ejemplo: el proceso de convertir los alimentos en energía.
2. *Regulación genética*: estas rutas son responsables de la regulación de los genes, y tienen una gran importancia porque los genes son responsables de la generación de proteínas, que a su vez son necesarios para cada una de las tareas de nuestro cuerpo.
3. *Transmisión de señales*: estas rutas permiten que la señal pase desde el exterior al interior de la célula y viceversa.

Los *Pathways* desempeñan un papel clave en los estudios de genómica avanzada, y es por ello, por lo que su inclusión en esta versión del modelo conceptual es necesaria. En este MCGH versión 2 se ha incluido el primero de los tres tipos de rutas biológicas, las "*rutas metabólicas*".

Se trata de una serie de reacciones químicas que conducen al sustrato inicial a uno o más productos finales. El producto final de una ruta metabólica puede usarse en tres formas diferentes: (1) para ser utilizado inmediatamente, (2) para iniciar una nueva ruta metabólica y (3) para ser almacenado en la célula.

La ruta metabólica está representada en el modelo conceptual como una combinación de eventos, representada por la relación entre los conceptos de "*Pathway*" y "*Event*", los cuales pueden ser de dos tipos. El primer tipo es un proceso atómico único, o, en otras palabras, para procesar el tipo más simple y no ser fiable a descomponerse en otras

más pequeñas (representados en el MC mediante la clase "*Process*"). El segundo tipo es un proceso complejo que consiste en una secuencia de otros procesos del tipo –*complejo o simple*-, representados en el MC mediante la clase "*Pathways*".

La asociación entre *Pathways* y *eventos* representa la composición de un "*Pathway*" (*ruta*) –*proporciona información sobre qué otros eventos anteriores forman parte de esta ruta metabólica*-. Para conocer el orden en la composición de los eventos en una ruta metabólica se define una relación reflexiva en la clase "*Event*". Las sustancias químicas que participan en un proceso están representadas en el MC mediante la clase "*Entity*", esta clase está relacionada con la clase "*Process*" por la clase "*Takes_part*", esto puede ocurrir en diferentes formas: (a) ser el químico principal, (b) como resultado del proceso y (c) ser un regulador del proceso –de dos tipos: activador o inhibidor-, representados en el modelo conceptual a través de las clases "*Input*", "*Output*" y "*Regulator*" respectivamente.

4.4 Descripción de Clases: MCGH v2

En este apartado se presenta la descripción de las clases definidas en la versión 2 del MCGH, con el objetivo de clarificar cada concepto representado en el modelo.

4.1.1 Vista Estructural

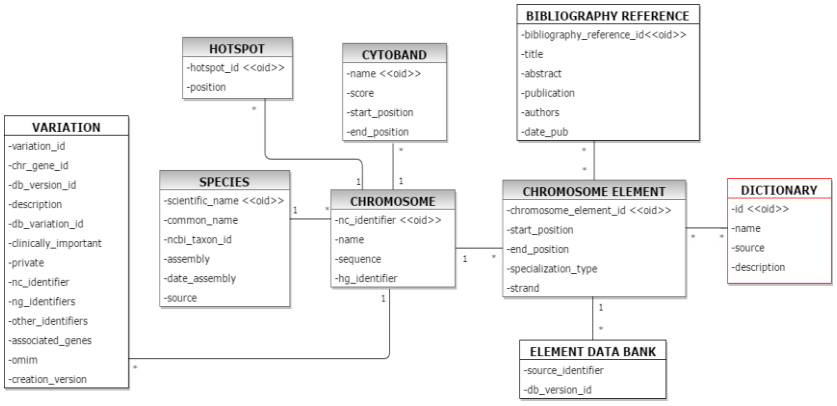


Figura 36. MCGH v2: "*Structural View*"

Esta vista, como su nombre indica, describe la estructura del genoma (Figura 36). A continuación, se describen cada una de las clases que forman esta vista:

Chromosome [CHROMOSOME] Es la clase principal de esta vista, y se define como una estructura organizada y única dentro del ADN (donde genes, elementos reguladores y otras secuencias de nucleótidos son localizados).		
Nombre	**Tipo dato**	**Descripción**
nc_identifier <<oid>>	String	Identificador interno de la secuencia cromosómica. Este identificador viene proporcionado por NCBI e identifica la secuencia del cromosoma, así como su versión (ej. NC_000013.10)
name	String	Los cromosomas se identifican por su nombre, que viene siendo un número según el orden en el que se encuentran dentro del ADN (ej. Ch1)
hg_identifier	String	Identificador de la versión del genoma de referencia utilizado (ej. HG_19) -*asocia cada cromosoma con la versión del genoma al que pertenece*-
sequence	Long	Secuencia de referencia del cromosoma (ej. AAGCTTCTCACCCTGTTCCTGCAT…)

Species Una especie se define como un grupo de organismos con ADN semejante. Esta clase sirve para determinar a qué familia pertenece cada uno de los cromosomas.		
Nombre	**Tipo dato**	**Descripción**
scientific_name <<oid>>	Short	Nombre científico e identificador por el cual se conoce la especie (ej. homo sapiens)
common_name	String	Nombre común por el cual se conoce la especie (ej. ser humano)
ncbi_taxon_id	String	Identificador dado a una especie por la organización de NCBI
assembly	String	Identificador de la versión utilizada como secuencia genómica de referencia de dicha especie
date_assembly	String	Fecha de la versión utilizada como secuencia genómica de referencia de dicha especie
source	String	Fuente de la cual se obtiene la secuencia genómica de referencia

Hotspot

Esta clase describe otra característica del cromosoma, los Hotspots representan información sobre los puntos en la secuencia de ADN donde existe mayor probabilidad de que se produzca la recombinación durante el proceso de meiosis (*los Hotspot representan los puntos de la secuencia donde se producen recombinaciones*).

Nombre	Tipo dato	Descripción
hotspot_id <<oid>>	String	Identificador interno del cruce de recombinación
position	-	Punto dentro de la secuencia de ADN en la que se produce el proceso de recombinación

Cytoband

Esta clase también conocida como "*banda citogenética*", describe otra característica del cromosoma representando información sobre las subregiones de un cromosoma que llegan a ser visibles microscópicamente después del tintado (tinción) durante una fase específica del ciclo celular. Son las zonas de un cromosoma que se diferencian con técnicas antiguas para localizar los genes mediante tinciones.

Nombre	Tipo dato	Descripción
name <<oid>>	String	La citobanda sigue siempre el mismo formato siguiendo las reglas establecidas, las cuales consisten en una "q" o una "p", dependiendo del brazo del cromosoma, seguida de uno, dos o tres números separados por puntos dependiendo de la resolución utilizada (ej. Q24.22)
score	String	Indica la intensidad de tintado, la cual puede tomar cinco valores diferentes proporcionales a la presencia de A y T
start_position	Long	Posición inicial en la secuencia de referencia del cromosoma
end_position	Long	Posición final en la secuencia de referencia del cromosoma

Chromosome element [CHR_ELEM]

Esta clase representa los fragmentos de un cromosoma que tienen alguna significación.

Nombre	Tipo dato	Descripción
chromosome_element_id <<oid>>	String	Identificador interno de cada uno de los elementos del cromosoma
start_position	String	Indica la posición inicial del elemento dentro del cromosoma
end_position	Long	Indica la posición final del elemento dentro del cromosoma

strand	Long	Hebra dentro de la doble hélice en la que se encuentra el elemento de cromosoma. Si un elemento se encuentra en la hebra positiva (*plus*) el elemento se lee de derecha a izquierda, en cambio si se trata de la hebra negativa (*minus*) el elemento se leerá en sentido contrario y además los nucleótidos serán invertidos
Specialization_type	String	Este atributo define el tipo de elemento que estamos definiendo. Estos pueden ser de tres tipos según nuestra descripción del dominio: *elementos transcribibles*, *elementos reguladores* y *regiones conservadas*

Los elementos del cromosoma pueden ser de tres tipos dependiendo de la función que desempeñen: (1) *elementos transcribibles*, (2) *elementos reguladores* y (3) *regiones conservadas*.

4.1.2 Vista de Transcripción

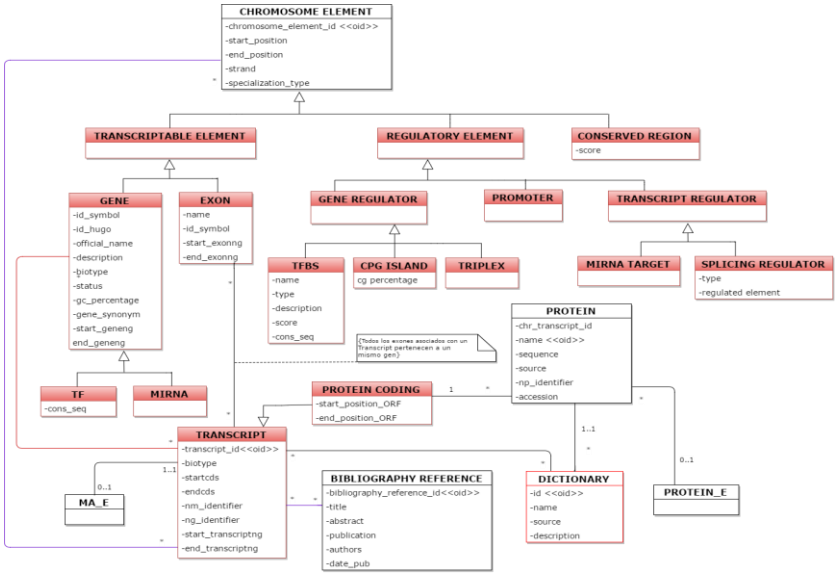


Figura 37. MCGH v2: "*Transcription View*"

Un gran número de genes expresan su funcionalidad a través de la producción de proteínas. La vista transcripción (Figura 37) muestra los componentes y conceptos relacionados con la síntesis de proteínas. A continuación, se describen las clases que forman la vista:

Transcriptable element
Esta clase representa una región del ADN que se puede transcribir, o en otras palabras un elemento del que se crea un ARN complementario a partir de la secuencia de ADN. Este tipo de regiones pueden especializarse en dos tipos: genes y exones. Estos son los elementos de cromosoma que se transcriben.

Gene [GENE]		
Esta clase representa una región de ADN que contiene la información necesaria para la síntesis de una macromolécula con una función celular específica, es decir contiene elementos reguladores que controlan el proceso de transcripción, normalmente sintetiza proteínas, pero también otro tipo de ARNs. Es la región del ADN con elementos reguladores (promotores, activadores, etc.) que contralan la transcripción.		
Nombre	**Tipo dato**	**Descripción**
id_symbol	String	Simbología utilizada para identificar un gen -*abreviatura unívoca con que la se identifica un gen*- (ej. BRCA1, NF1, APC, FBN1)
id_hugo	String	Valor único y significativo de los genes que viene dado por el consorcio HGNC (*HUGO Gene Nomenclature Committee*), como, por ejemplo: para BRCA2 su identificador HUGO sería '1101'
official_name	String	Nombre oficial y completo asignado a un gen (ej. BRCA1: Homo sapiens breast cancer 1, early onset)
description	String	Descripción del gen (resumen sobre su funcionalidad) al que se hace referencia
biotype	String	Especialización del tipo de gen dependiendo de las funciones que realiza, puede tomar valores como, por ejemplo: snRNA, miRNA, protein coding, etc.
status	String	Determina el estado de validez en el que se encuentra cada elemento en la actualidad (ej. Validados, en periodo de estudio)
gc_percentage	Float	A diferencia del resto de regiones de la secuencia de ADN, ha sido comprobado que las regiones transcribibles tienen mayor alto contenido de Gs y Cs en su secuencia y que dicho contenido es directamente proporcional a la longitud de la secuencia

		codificante. Este atributo almacena el porcentaje de pares de bases Cs y Gs que existen en el elemento
gene_synonym	String	Sinónimos reconocidos para una simbología de un gen (ej. BRCA1 tiene como genes sinónimos: BRCAI; BRCC1; BROVCA1; IRIS; PNCA4; PPP1R53; PSCP; RNF53)
start_geneng	Int	Posición de inicio del gen en la secuencia –a nivel génico- (ej. '92501')
end_geneng	Int	Posición del final del gen en la secuencia –a nivel génico- (ej. '173689')

Un gen, dependiendo del valor de su atributo "*biotype*" puede especializarse en diversos tipos de genes, dependiendo como se ha dicho anteriormente de las funciones que desempeñe. Existen muchos tipos de genes que podrían ser modelados, pero por simplificar el modelo se decide ilustrar un solo ejemplo, los factores de transcripción que se describen a continuación.

Tf

La clase "tf" (*factor de transcripción*) representa aquellos genes que codifican una proteína cuya función es regular la transcripción de otros genes o incluyo la suya propia. Son genes que producen factores de transcripción.

Nombre	Tipo dato	Descripción
cons_seq	-	Hace referencia a la secuencia de nucleótidos que una vez acoplada a las regiones de unión de la cadena de ADN realizará una función reguladora para el gen (es una secuencia consenso)

miRNA

Esta clase representa los MicroRNA. Los Micro-RNA (miRNA) son una clase de RNA pequeños no codificantes (aquellos que no codifican para proteínas) que regulan la expresión génica pos-transcripcional [119].

Exon

Esta clase representa un elemento transcribible que forma parte del gen, y que es además la unidad básica de los transcritos. Cada exón codifica una porción específica de la proteína completa, de manera que el conjunto de exones forma la región codificante del gen.

Nombre	Tipo dato	Descripción
name	String	Nombre e identificador del exón -

		nombre que proporciona NCBI a cada uno de los exones que forman parte de un gen- (ej. 1a, 4b, 10a)
id_symbol	String	Simbología utilizada para identificar un gen, clave ajena a la tabla Gene que ayuda a conocer qué exones forman parte de cada gen (ej. ATP7B, NF2, COL1A2)
start_exonng	Int	Posición de inicio del exón en la secuencia de referencia del gen (ej. '92501')
end_exonng	Int	Posición donde finaliza el exón en la secuencia de referencia del gen (ej. '92713')

Regulatory element

Esta clase representa regiones del ADN que realizan una función reguladora controlando ciertos procesos existentes dentro el ADN. Los elementos reguladores se especializan en dos clases dependiendo de si es un elemento regulador del gen o del transcrito: "*gene regulator*" y "*transcript regulator*".

Gene regulator

Esta clase representa los elementos reguladores del gen, entre los cuales se encuentran: *tfbs*, *cpg_island* y *triplex*.

Tfbs

La clase "*tfbs*" (transcription factor binding sites) son regiones de unión de los factores de transcripción que producen un efecto en la transcripción del gen bien sea de activación o represión. TF: proteína.

Nombre	Tipo dato	Descripción
name	Long	Nombre que toma el sitio de unión de los factores de transcripción
type	String	Los sitios de unión de los factores de transcripción pueden ser de dos tipos dependiendo de la función que desempeñen: *activador* o *inhibidor*
description	String	Descripción del tfbs
score	Float	Grado de similitud entre la secuencia consenso y el tfbs
cons_seq	-	Secuencia consenso la cual enlaza el tfbs

Cpg island

Las "*cpg island*" conforman aproximadamente un 40% de promotores de los genes de mamíferos. Son regiones donde existe una gran concentración de pares de Cs y Gs enlazados por fosfatos. La "p" en CpG representa que están enlazados por un fosfato y simboliza un conjunto de repeticiones de las bases CG que están cerca del promotor y son objetivos para la metilación que es otra manera de alterar la expresión del gen. La definición formal de una "*CpG island*" es una región con al menos 200 pares de bases, con un porcentaje de GC mayor de 50 y con un promedio de CpG observado/esperado mayor de 0,6.

Nombre	Tipo dato	Descripción
cg_percentage	Float	Representa el porcentaje de GC en el elemento

Triplex

Los "*triplex*" son secuencias de ADN que se intercalan en la doble hélice de ADN de las células, pasando este a tener tres cadenas, de tal manera que se impide el proceso de transcripción causando un efecto negativo en el individuo (*zonas con triple hélice*).

Promoter

Esta clase representa la región de ADN que controla la iniciación de la transcripción de una determinada parte del ADN a ARN (determina el lugar donde la ARN polimerasa comienza la transcripción de un gen).

Transcript regulator

Esta clase representa regiones reguladoras del transcrito. Existen muchas especializaciones de elementos reguladores del transcrito, pero por razones de simplificación en este modelo se representan únicamente dos: "*mirna target*" y "*splicing regulator*" (*elemento regulador de la transcripción*).

Mirna target

Esta clase representa una región reguladora del transcrito a la que se unirá post-transcripcionalmente un miRNA (*dianas de los miRNA*).

Splicing regulator

Esta clase representa un elemento regulador del transcrito que regula el proceso de *splicing*.

Nombre	Tipo dato	Descripción
type	-	Indica el tipo de regulación y puede tomar dos valores, desactivar (*silencer*) o promover (*enhancer*)
regulated_element	-	Indica cual es el elemento regulado si se trata de un intrón o un exón

Conserved region Esta clase representa las regiones conservadas dentro del cromosoma, regiones que normalmente tienden a ser no codificantes, es decir, se mantienen intactas tras el proceso de evolución entre las especies.		
Nombre	**Tipo dato**	**Descripción**
score	Float	Grado de conservación de la región, (o un valor estadístico que indica una probabilidad o un valor proveniente de una formula), un número real, cuanto más grande más conservado

Transcript [TRANSCRIPT] Esta clase representa los diferentes transcritos que presenta un gen. Estos transcritos están formados por una serie de exones. Como se ha comentado antes, existe un fenómeno llamado "*splicing alternativo*" que permite la combinación de diferentes exones, e incluso en poca medida algún intrón, formando diferentes transcritos.		
Nombre	**Tipo dato**	**Descripción**
transcript_id <<oid>>	Int	Identificador interno del transcrito
biotype	String	Cada transcrito puede tener una función diferente representada con este atributo, que puede tomar el valor de: *Protein Coding*, *tRNA*, *rRNA*, *miRNA*, *siRNA*, *piRNA*, *Antisense*, *Long noncoding*, *Riboswich*, *snRNA*, *snoRNA*, *mitocondrial*
startcds	Int	Define cual es la posición inicial del proceso de traducción a proteína. Este proceso empieza siempre por el codón de inicio de la traducción que es "ATG", aunque no tiene por qué ser el primero
endcds	Int	Define cual es la posición final del proceso de traducción
nm_identifier	String	Identificador que proporciona NCBI a la secuencia del transcrito, así como su versión
ng_identifier	String	Clave ajena a la tabla "*Sequence_NG*" que asocia a que gen pertenece este transcrito
start_transcriptng	Int	Define cual es el inicio del transcrito dentro de la secuencia del gen

end_transcriptng	Int	Define cual es el final del transcrito en la secuencia del gen

La relación entre la clase "*Transcript*" y "*Exon*" (*perteneciente a la vista estructural*) indica qué exones forman cada uno de los transcritos.

Protein coding Esta clase es una especialización de la clase transcrito y que, como su propio nombre indica, representa el primero de los biotipos anteriormente mencionado. Este tipo de transcritos sintetiza para una proteína.		
Nombre	**Tipo dato**	**Descripción**
start_position_ORF	-	Indica la posición de inicio de la secuencia codificada
end_position_ORF	-	Indica la posición final de la secuencia codifica

Protein Esta clase da soporte a los miles de proteínas que se sintetizan a partir de un transcrito. Algunos ejemplos de bases de datos especializadas en "*Proteínas*" están: *UniProt*, *Swissprot*, *TrEMBL*, etc.		
Nombre	**Tipo dato**	**Descripción**
chr_transcript_id	Int	Identificador interno del transcrito (clave ajena a la tabla "Transcript" e identificador del transcrito)
name <<oid>>	String	Nombre e identificador de la proteína
sequence	Long	La secuencia de la proteína
source	String	Fuente de datos de la cual se ha extraído la información (ej. NCBI)
np_identifier	String	Identificador interno que proporciona NCBI acerca de la proteína, así como su versión. (ej. NP_000681.2:p.Glu504Lys)
accession*	String	Identificador que presenta la proteína en la fuente de datos de la cual ha sido extraída

4.1.3 Vista de Variaciones

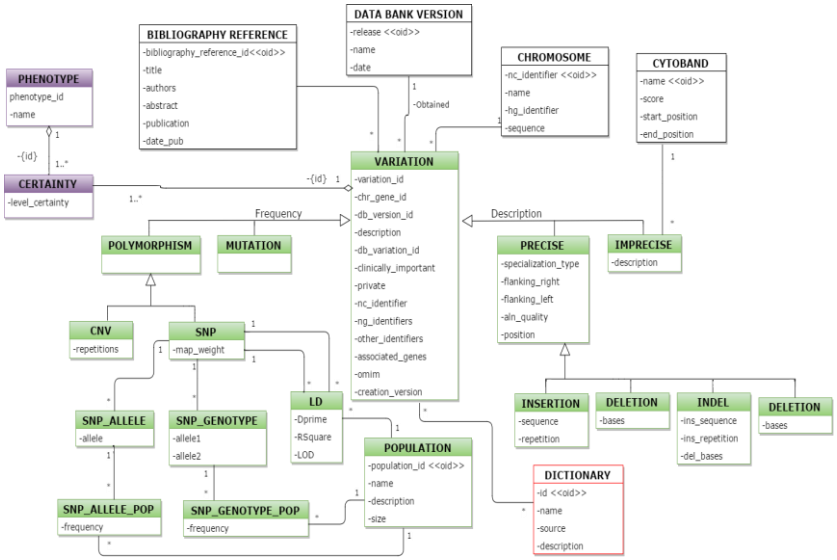


Figura 38. MCGH v2: "*Variation View*"

La vista variación modela el conocimiento relacionado con las diferencias encontradas en la secuencia de ADN de diversos individuos (Figura 38). A continuación, se detallan las clases que la forman y la explicación de cada una de ellas:

Variation [VARIATION] Esta es la clase principal en esta vista, en ella se representan como su propio nombre indica, todas las variaciones existentes en la cadena de ADN. Cambios en la secuencia respecto a la secuencia de referencia.		
Nombre	**Tipo dato**	**Descripción**
variation_id <<oid>>	Int	Nombre e identificador de la variación
chr_gene_id	Int	Clave ajena a la tabla Gene
db_version_id	Int	Indica a qué versión y de qué base de datos se ha extraído dicha variación
description	String	Proporciona una descripción de la variación
db_variation_id	String	Identificador que proporciona la fuente de datos de la cual se ha extraído la variación

clinically_important	String	Define la importancia clínica dada por la fuente de datos, como, por ejemplo: no probada, no-patogénica, etc.
private	Int	este es un booleano rellenado por decisión del usuario. Aquí especificamos si la variación es de tipo privada o no
nc_identifier	String	Versión de la secuencia cromosómica respecto a la que viene dada la variación (ha de ser el mismo que el que tenemos ya cargado)
ng_identifier	String	Versión de la secuencia génica respecto a la que viene dada la variación (guardamos de cualquier versión)
other_identifiers	String	Otros posibles identificadores que presenten las fuentes de datos (ej. nt´s, nr´s, etcétera)
associated_genes	String	Genes asociados a la variación
omim	String	Versión de OMIM (*Online Mendelian Inheritance in Man*) para la variación
creation_version	String	Indica la versión de la variación (al ser creada)

Las variaciones se especializan siguiendo dos criterios: la precisión en su descripción (ISA *description*) y su frecuencia (ISA *frequency*). En la jerarquía frecuencia, o en otras palabras si la variación se presenta en más del 1% de la población o es un caso puntual, una variación puede estar especializada en dos clases: "*Mutation*" y "*Polimorphysm*". En la jerarquía descripción, una variación puede estar especializada en dos clases: "*Precise*" e "*Imprecise*", dependiendo de si se conocen datos al respecto de su posición.

Por otra parte, cabe destacar que la clase "*Variation*" enlaza esta vista con la vista estructural del genoma, mediante una relación entre la clase "*Variation*" y la clase "*Chromosome_element*" que indica que una variación es un elemento que forma parte de un cromosoma.

Mutation
Esta clase es una especialización de tipo *ISA frequency*, hace referencia a las variaciones con efecto patológico que se encuentran en un bajo porcentaje de la población, es decir, en menos del 1%.

Polymorphism
Esta clase es una especialización de tipo *ISA frequency*, describe las variaciones que aparecen en más del 1% de la población y normalmente no tienen un diagnostico maligno, por lo que se heredan de generación en generación. Este tipo de variaciones puede especializarse en dos tipos: "*CNV*" (*Copy Number Variation*) y "*SNP*" (*Single Nucleotide Polymorphism*).

CNV		
Un "*CNV*" (*copy number variation*) es definido como una variación que consiste en la repetición un cierto número de veces o el borrado de una pequeña región de la secuencia de ADN.		
Nombre	**Tipo dato**	**Descripción**
repetitions	-	Atributo multivaluado con las repeticiones habituales de ese elemento (CNP). Este atributo almacena el número de veces que la secuencia se repite o se borra

SNP		
Un "*SNP*" es un polimorfismo que tiene lugar cuando un único nucleótido dentro del genoma difiere de lo habitual entre individuos de la misma especie agrupados por poblaciones. Constituyen hasta el 90% de todas las variaciones genómicas humanas, y aparecen cada 1300 bases en promedio a lo largo del genoma humano.		
Nombre	**Tipo dato**	**Descripción**
map_weight	Int	Es el número de veces que se localiza el SNP en el genoma (las veces que se ha mapeado el SNP en la muestra del genoma de un individuo)

Un SNP es un cambio de un único nucleótido en una posición del genoma, pero a su vez puede proporcionar datos relevantes: los distintos valores que puede tomar el SNP teniendo en cuenta un único alelo ("*SNP_Allele*") y las diferentes combinaciones de valores que puede tomar el SNP teniendo en cuenta los dos alelos ("*SNP_Genotype*").

Además, existe más información de interés con respecto a los SNPs, así como el "*Linkage Disequilibrium*" (LD) y que se describe como marcador que indica la relación existente entre dos SNPs dentro de una población.

SNP_Allele		
Esta clase representa los diferentes valores que puede tomar un SNP teniendo en cuenta un solo alelo -*Los diferentes alelos que pueden aparecer asociados al SNP (bases diferentes que aparecen)*-.		
Nombre	**Tipo dato**	**Descripción**
allele	Char	Este atributo indica el valor que puede tomar el alelo en cada caso. Su dominio es {A,T,G,C}

SNP_Genotype		
Esta clase representa los diferentes valores que pueden tomar el par de alelos de cada individuo en la posición del SNP teniendo en cuenta las dos hebras -*diferentes combinaciones de alelos teniendo en cuenta las dos hebras*-.		
Nombre	**Tipo dato**	**Descripción**
allele1	String	Este atributo indica el valor que puede tomar el alelo en una hebra. Su dominio es {A,T,G,C}
allele2	String	Este atributo indica el valor que puede tomar el alelo en la otra hebra. Su dominio es {A,T,G,C}

Como se ha comentado en la descripción de SNP, cada uno de ellos está directamente relacionado con varias poblaciones, por lo que las dos clases, "*SNP_Allele*" y "*SNP_Genotype*" tienen relación con varias poblaciones en cada caso.

Para proporcionar información sobre la frecuencia de aparición de cada SNP en diferentes poblaciones, bien sea a nivel *alélico* o a nivel *genotípico*, se crean también las clases "*SNP_Allele_Pop*" y "*SNP_Genotype_Pop*".

SNP_Allele-Pop		
Esta clase representa la frecuencia en la que cada SNP, teniendo únicamente en cuenta un alelo, aparece en cada población.		
Nombre	**Tipo dato**	**Descripción**
frequency	Float	Frecuencia con la que cada SNP aparece en diversas poblaciones (valor absoluto de veces que se da la variación)

SNP_Genotype-Pop Esta clase representa la frecuencia en la que cada SNP aparece en cada población teniendo únicamente en cuenta los dos alelos.		
Nombre	**Tipo dato**	**Descripción**
frequency	Float	Frecuencia con la que cada SNP aparece en diversas poblaciones

Population Esta clase representa conjuntos de individuos con características comunes -*grupo de individuos sobre los que han hecho un estudio-*.		
Nombre	**Tipo dato**	**Descripción**
id	-	Identificador interno de la población
name	-	Nombre e identificador de cada población
description	-	Descripción de cada población
size	-	Cantidad de individuos pertenecientes a una población

LD Otro concepto modelado que hemos mencionado anteriormente es el "*Linkage Disequilibrium*" o "*LD*", que es un marcador que define la relación existente entre dos SNPs en una población específica.		
Nombre	**Tipo dato**	**Descripción**
Dprime	Double	Los tres son valores matemáticos de ámbito muy biológico los cuales vamos a explicar con mayor detalle más adelante
Rsquare	Double	
LOD	Double	

Precise [PRECISE] Esta clase representa las variaciones detectadas con posición conocida dentro del cromosoma en la secuencia de ADN (variación cuyos datos permiten ubicarla con precisión en la secuencia del cromosoma).		
Nombre	**Tipo dato**	**Descripción**
specialization_type	String	Indica el tipo de variación tuvo lugar dentro del genoma (ej. ID –*indel-*, DE –*deletion-*, etc.)
flanking_right	String	Presenta una secuencia de 20 nucleótidos que se encuentran a la derecha de la variación
flanking_left	String	Presenta una secuencia de 20 nucleótidos que se encuentran a la

		izquierda de la variación
aln_quality	Int	Indica la calidad del alineamiento dentro del gen
position	Int	Posición en la que se encuentra la variación dentro de la secuencia del cromosoma

La clase "*Precise*" se especializa en cuatro nuevas entidades dependiendo de qué tipo de variación haya tenido lugar dentro del genoma: "*Insertion*", "*Deletion*", "*Indel*" e "*Inversion*".

Insertion		
Esta clase representa variaciones que consisten en la inserción de una secuencia de nucleótidos un número de veces en la secuencia de ADN del cromosoma.		
Nombre	**Tipo dato**	**Descripción**
sequence	String	Secuencia de nucleótidos insertados en la secuencia
repetition	Int	Número de veces que se repite la secuencia insertada

Deletion		
Esta clase representa variaciones que consisten en el borrado de un número de nucleótidos en la secuencia de ADN del cromosoma.		
Nombre	**Tipo dato**	**Descripción**
bases	Int	Número de nucleótidos borrados en la secuencia

Indel		
Esta clase representa variaciones consistentes en inserciones y borrados a la vez en la secuencia de ADN del cromosoma.		
Nombre	**Tipo dato**	**Descripción**
ins_sequence	String	Secuencia de nucleótidos insertados en la secuencia
ins_repetition	Int	Número de veces que se repite la secuencia insertada
del_bases	Int	Número de nucleótidos borrados

Inversion Esta clase representa variaciones que invierten el orden de una secuencia de nucleótidos en la secuencia del cromosoma.		
Nombre	**Tipo dato**	**Descripción**
bases	Int	Número de nucleótidos invertidos en la secuencia

Imprecise [IMPRECISE] La clase "*Imprecise*" dentro de la jerarquía de descripción representa variaciones cuya posición es desconocida dentro de la secuencia de ADN.		
Nombre	**Tipo dato**	**Descripción**
description	String	Descripción de la variación en lenguaje natural

En esta versión del modelo se realiza un cambio en la representación de "*Fenotipos*", con el objetivo de gestionar este conocimiento de una manera más simplista y clara (como se aprecia en la figura 39).

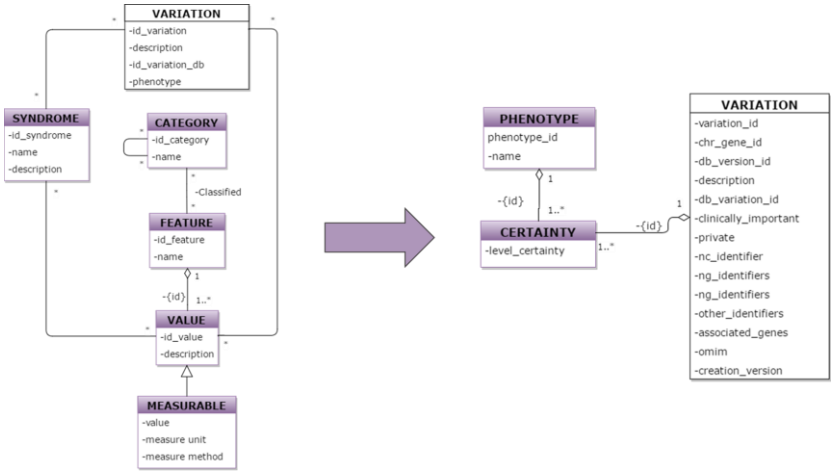


Figura 39. "*Phenotype View*": Desde versión 1.1 a versión 2 del MCGH

Phenotype [PHENOTYPE] Esta clase representa los fenotipos asociados a una o varias variaciones del ADN. El fenotipo es el conjunto de caracteres visibles que un individuo presenta como resultado de la interacción entre su genotipo y el medio.		
Nombre	**Tipo dato**	**Descripción**
phenotype_id	Int	Identificador interno del fenotipo

name	String	Nombre del fenotipo

Certainty [CERTAINTY] Esta clase representa los niveles de certeza (certidumbre) asociados a ciertos fenotipos con respecto a una o varias variaciones.		
Nombre	**Tipo dato**	**Descripción**
level_certainty	String	Representa el nivel de certeza entre fenotipo y variación (nivel de certeza que facilitan los estudios -*fuentes*- sobre las variaciones)

4.1.4 Vista de Rutas Metabólicas

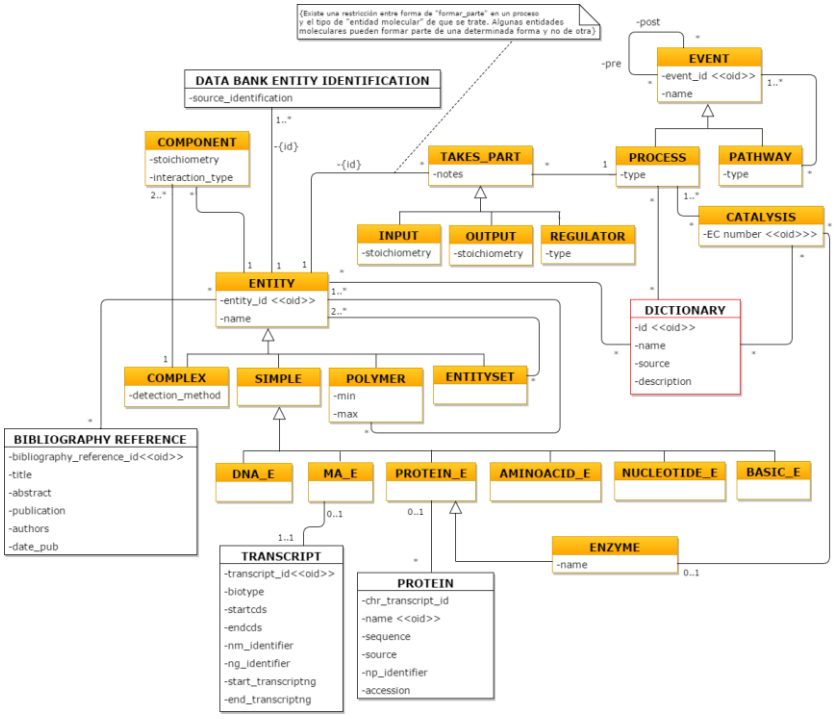


Figura 40. MCGH v2: "*Pathway View*"

En bioquímica, las rutas metabólicas (*Pathways*), son una serie de reacciones químicas que ocurren dentro de una célula (Figura 40). Esta composición de procesos viene representada en el modelo por las siguientes clases:

Event Esta es la clase principal y es la que representa la combinación de procesos existentes en el organismo.		
Nombre	**Tipo dato**	**Descripción**
event_id <<oid>>	String	Identificador interno del evento
name	-	Nombre que tiene el evento

Además, la clase "*Event*" se especializa en dos clases dependiendo de la cantidad de procesos que lo formen: "*Process*" y "*Pathway*".

Process Esta clase representa un único proceso atómico o dicho en otras palabras un proceso de tipo simple.		
Nombre	**Tipo dato**	**Descripción**
type	String	Especifica qué tipo de proceso se ha llevado a cabo

Pathway Esta clase representa un proceso complejo formado por una secuencia de otros procesos de tipo complejo o simple (*ruta metabólica compuesta de procesos o de otras rutas más simples*).		
Nombre	**Tipo dato**	**Descripción**
type	String	Define el tipo de *Pathway* activado (ejecutado)

La asociación entre las clases "*Pathway*" y "*Event*" representa la composición de un *Pathway*, es decir, proporciona información sobre que otros eventos anteriores forman parte de dicho *Pathway*. Por otra parte, la relación existente de la clase "*Event*" consigo misma cuyas aristas poseen el nombre de "*Pre*" y "*Post*" nos permite conocer el orden de la composición de los eventos dentro del *Pathway*. El que va primero toma el valor de Pre y el que lo sigue toma el valor de Post.

Por otra parte, y siguiendo con la descripción de la vista, una entidad puede participar en un proceso de varias maneras:

a) *Siendo el químico principal*, es decir la entrada necesaria para ese proceso, a veces también llamada sustrato.

b) *Como resultado del proceso* o en otras palabras la salida o producto final, o

c) *Siendo un regulador del proceso*, de los cuales se pueden distinguir dos tipos: *activador* e *inhibidor*. Por otro lado, existe

un tipo especial de elemento regulador, la "*catálisis*", de la que a veces se desconoce información al respecto, pero es sabida su existencia en algunos procesos. Este conocimiento se modela mediante las siguientes clases: "*Takes_part*", "*Input*", "*Output*" y "*Regulator*".

Takes_part Es una clase genérica que define de qué manera una entidad participa dentro de uno o varios procesos (*Representa la participación de una entidad en un proceso*).		
Nombre	**Tipo dato**	**Descripción**
notes	String	Comentario sobre la relación entre las entidades que toman parte en cada proceso

Se especializa en tres entidades diferentes dependiendo de la manera en la que dicha entidad participe en el proceso: "*Input*", "*Output*" y "*Regulator*".

Input Esta clase representa la entidad de entrada a un proceso.		
Nombre	**Tipo dato**	**Descripción**
stoichiometry	Int	Cantidad de la entidad que interviene en el proceso. Cantidad usada de la entidad entrante en el proceso

Output Esta clase representa el resultado final del proceso.		
Nombre	**Tipo dato**	**Descripción**
stoichiometry	Int	Cantidad producida de la entidad saliente por el proceso

Regulator Esta clase como su propio nombre indica los procesos reguladores existentes en las partes intermedias de la reacción *-entidad que controla un proceso: activándola o inhibiéndola-*.		
Nombre	**Tipo dato**	**Descripción**
type	-	Entidad que controla un proceso "*activándola*" o "*inhibiéndola*"

Catalysis		
Esta clase define el proceso por el cual se aumenta o disminuye la velocidad de una reacción química. Es un tipo especial de regulador de *Pathways* que ha sido modelada aparte debido al hecho de que se tiene constancia de que forma parte de muchos procesos, pero en algunos de ellos el catalizador es desconocido. En los casos en los que el catalizador es conocido, una enzima es asociada al correspondiente proceso. Reacción catalizadora.		
Nombre	**Tipo dato**	**Descripción**
EC number <<oid>>	-	Identificador de la reacción asignado por la *Enzyme Commision*

Los números EC (*Enzyme Commission numbers*) son un esquema de clasificación numérica para las enzimas, basado en las reacciones químicas que catalizan.

En realidad, los números EC codifican reacciones catalizadas por enzimas. Enzimas diferentes (por ejemplo, que procedan de organismos diferentes) que catalicen la misma reacción recibirán el mismo número EC. Cada código de enzimas consiste en las dos letras EC seguidas por 4 números separados por puntos. Estos números representan una clasificación progresivamente más específica. Por ejemplo, la "*enzima tripéptido aminopeptidasa*" tiene el código EC 3.4.11.4.

Enzyme		
Esta clase es una especialización de "*Proteína*" que cataliza reacciones químicas. Una enzima hace que una reacción química que es energéticamente posible pero que transcurre a una velocidad muy baja, sea cinéticamente favorable, es decir, transcurra a mayor velocidad que sin la presencia de la enzima. Está asociada con el proceso de catálisis para determinar cuál es el catalizador en caso de ser conocido.		
Nombre	**Tipo dato**	**Descripción**
name	-	Las enzimas son usualmente nombradas de acuerdo con la reacción que producen. Normalmente, el sufijo "*-asa*" es agregado al nombre del sustrato (por ejemplo, la lactasa es la enzima que degrada lactosa) o al tipo de reacción (por ejemplo, el ADN polimerasa forma polímeros de ADN)

Entity Esta es la clase genérica que representa el tipo de entidades que pueden participar en un proceso de un *Pathway*. Una entidad que toma parte en un proceso biológico.		
Nombre	**Tipo dato**	**Descripción**
entity_id <<oid>>	String	Identificador interno de la clase "*Entity*"
name	String	Atributo genérico que proporciona información acerca del nombre de la entidad

La clase "*Entity*" se especializa en cuatro clases dependiendo del tipo de entidad: "*Complex*", "*Polymer*", "*Simple*" y "*EntitySet*".

Complex Esta clase representa entidades que están formadas por la combinación de otras entidades más simples. Es una entidad compleja, compuesta de 2 o más entidades.		
Nombre	**Tipo dato**	**Descripción**
detection_method	-	Este atributo indica la técnica usada para determinar cómo se ha formado la entidad

Component Esta clase representa de qué manera una entidad "*Complex*" está formada por sus entidades más simples.		
Nombre	**Tipo dato**	**Descripción**
stoichiometry	Int	Cantidad en la que una entidad participa en una entidad compleja
interaction_type	-	Permite conocer como la -*entidad compleja*- ha sido formado a partir de cada uno de sus componentes

Polymer Esta clase representa entidades que son generadas por la repetición de alguna entidad, bien sea compleja o simple. Es el compuesto que se forma por repetición de una entidad simple o compuesta.		
Nombre	**Tipo dato**	**Descripción**
min	-	Representa el rango de repeticiones mínimo de la entidad que forma el polímero
max	-	Representa el rango de repeticiones

		máximo de la entidad que forma el polímero

Simple

Esta clase representa las entidades más simples que pueden formar parte de un proceso, como por ejemplo: "*gen*", "*ARN*", "*proteína*", "*aminoácido*", "*nucleótido*", "*entidad básica*" (agua, fósforo, etc.), representadas en el modelo conceptual a través de una especialización con la clase "*Simple*" y las clases "*dna_e*", "*rna_e*", "*protein_e*", "*aminoacid_e*", "*nucleotide_e*" y "*basic_e*" respectivamente.

EntitySet

Esta clase representa un conjunto de entidades que participan de manera habitual conjuntamente en algunos procesos, lo que permite reducir la cantidad de procesos similares existentes. Entidades funcionalmente equivalentes que toman parte de la misma forma en un proceso.

4.1.5 Vista de Fuentes de Datos y Bibliografía

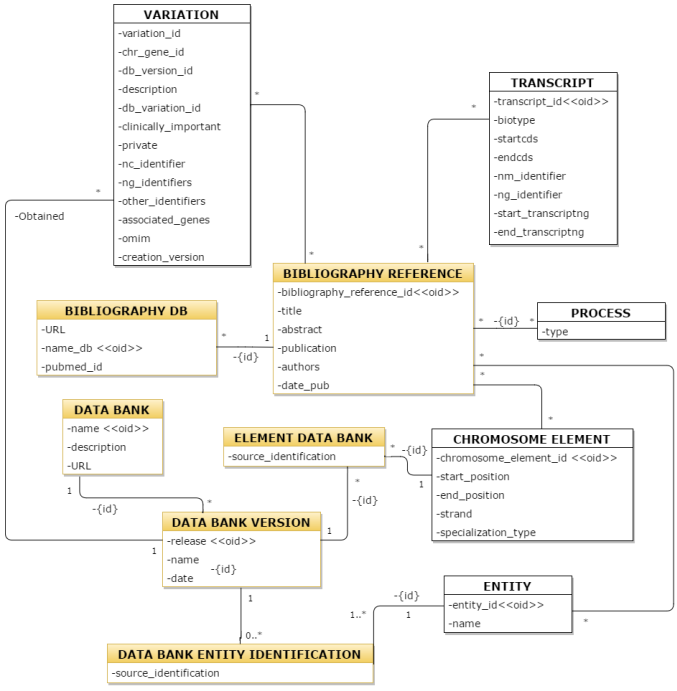


Figura 41. MCGH v2: "*Bibliography and data bank View*"

Esta vista proporciona información sobre las fuentes de datos de las que se ha extraído la información que se va a almacenar en el modelo (Figura 41), así como una serie de documentos bibliográficos de consulta para quien desee obtener más información con respecto a algún aspecto aquí definido.

Para mantener información sobre las fuentes de las cuales se ha obtenido la información, esta vista presenta las siguientes clases:

Data bank [DATABANK] Esta clase proporciona información sobre la fuente de datos de la cual se extrae la información de cada uno de los elementos del modelo.		
Nombre	**Tipo dato**	**Descripción**
name <<oid>>	String	Nombre de la fuente de datos (ej. NCBI)
description	String	Descripción de la fuente de datos (ej. National Center for Biotechnology Information)
url	String	Dirección web (enlace) de la información

Data bank version Esta clase proporciona información sobre la versión de cada una de las bases de datos que se han utilizado y en qué fecha dichas bases de datos han sido actualizadas.		
Nombre	**Tipo dato**	**Descripción**
release <<oid>>	String	Versión de la fuente de datos
name	String	Nombre de la fuente de datos utilizada
date	Date	Fecha en la que se actualizó por última vez la fuente de datos consultada

La clase "*Variation*" se relaciona con esta clase para obtener información acerca de la proveniencia de sus datos.

Element data bank [ELEMENT_DATABANK] Esta clase permite relacionar cada uno de los elementos del cromosoma de que fuente de datos han sido extraídos y su versión.		
Nombre	**Tipo dato**	**Descripción**
db_version_id	Int	clave ajena a la tabla "*Databank_Version*" que indica a

		que versión de qué base de datos se asocia cada elemento (Identificador de la versión de la fuente de datos)
source_identifier	String	Identificador del elemento de cromosoma en el banco de datos. Este atributo indica el identificador que proporciona cada una de las fuentes a los elementos del cromosoma

Data Bank Entity Identification		
Esta clase permite relacionar cada una de las entidades que forman los *Pathways* con la fuente de datos y la versión de la cual se ha extraído la información.		
Nombre	**Tipo dato**	**Descripción**
source_identification	String	Este atributo indica el identificador que proporciona cada una de las fuentes a las entidades que forman los *Pathways*

Por último, para mantener información sobre la bibliografía asociada a cada elemento, la vista incluye también las siguientes clases:

Bibliography DB [BIBLIOGRAPHY_DB]		
Esta clase representa las distintas fuentes de datos de la web de las que se extraen las publicaciones científicas.		
Nombre	**Tipo dato**	**Descripción**
url	String	Dirección web de la base de datos de las que se extraen las publicaciones
name_db <<oid>>	String	Nombre de la base de datos de la que se extraen las publicaciones científicas (ej. Pubmed)
pubmed_id	Int	Identificador que la base de datos de Pubmed proporciona al artículo

Bibliography reference [BIB-REF] Esta clase proporciona información sobre los artículos relacionados con cada uno de los elementos almacenados si se dispone de ella.		
Nombre	**Tipo dato**	**Descripción**
bibliography_reference_id <<oid>>	Int	Identificador interno de las referencias bibliográficas
title	String	Título del artículo
abstract	String	Resumen del artículo
publication	String	Contiene la referencia del artículo (ej. Langston et al., New Engl J Med 334:137, 1996)
authors	String	Autores que han escrito el artículo
date_pub	String	Fecha en la cual se ha publicado el artículo

Las clases "*Variation*", "*Process*", "*Transcript*", "*Chromosome element*" y "*Entity*" se relacionan con esta clase.

4.5 Conclusiones

En este capítulo se ha planteado un *Modelo Conceptual del Genoma Humano* (MCGH) con el objetivo de facilitar una definición clara (*conceptual*) del comportamiento del genoma humano. La complejidad del dominio genómico requiere de la aplicación de técnicas de *Ingeniería de Software* (IS) que ayuden a definir la amplia cantidad de factores, elementos o conceptos participantes en el funcionamiento de la vida.

En este trabajo se han aplicado técnicas de *modelado conceptual*, logrando generar una representación holística del dominio. Esta representación conceptual fue desarrollada junto a expertos de las áreas de *modelado conceptual*[18], *biología molecular*[19] (por ejemplo, biólogos, médicos especializados en genética, entre otras) y *colaboradores* del *grupo genoma* del Centro PROS. Mediante este modelo se logró un mejor entendimiento del genoma, lo cual permitió evolucionar de una versión a otra (v1 a v2) aportando las justificaciones asociadas a cada decisión tomada en el proceso de crecimiento. Estas decisiones permitieron generar una definición integral del genoma *-formal, rigurosa y extensible-*, la cual cubre las necesidades detectadas a causa de la evolución natural del dominio.

Para cerrar este capítulo se presentó una descripción de todas las clases definidas en la versión 2 del MCGH. Esto con el fin de definir los conceptos participantes en el modelo (*nombre*, *tipo de dato* y *descripción*). En este contexto se justifica el principal resultado de esta Tesis Doctoral: un Modelo Conceptual holístico del Genoma Humano como herramienta conceptual esencial, el cual permite disponer de plataformas que minimicen el problema derivado del caos de datos genómico en el que se encuentra la Bioinformática actual.
Este modelo conceptual será la base para el desarrollo de un prototipo que permita gestionar los datos genómicos facilitados por los distintos repositorios, pero antes de presentar ese prototipo, el próximo capítulo tiene objetivo presentar la versatilidad del modelo a través de la inclusión de nuevo conocimiento conservando su definición inicial, para de esta forma apoyar la solución del tercer subobjetivo planteado en esta Tesis Doctoral.

---

[18] *Dr. Óscar Pastor López, Dr. Juan Carlos Casamayor Ródenas, Dra. Laura Mota, Dra. Matilde Celma Giménez & Dra. M. Ángeles Pastor.*

[19] *Dra. Ana M. Levin & Centro de Investigación Príncipe Felipe (CIPF)*

CAPÍTULO 5

Estrategia de Integración de Haplotipos al MCGH

La evolución constante en el dominio genómico contribuye cada día a la generación de grandes cantidades de datos nuevos, lo que significa que, si no se gestiona correctamente, la calidad de los datos podría verse comprometida (como, por ejemplo, problemas relacionados con la *heterogeneidad* e *inconsistencia* de los datos).

En este capítulo se propone utilizar el *Modelo Conceptual del Genoma Humano* (MCGH) explicado en el capítulo anterior, con el objetivo de comprender y mejorar el compromiso ontológico con el dominio –*genómico*- y de esta forma extender el MCGH con la integración de un nuevo concepto: "*Haplotipos*".

El objetivo es mejorar la comprensión de la relación entre el "*genotipo*" y "*fenotipo*", debido a que los nuevos hallazgos demuestran que este caso es más complejo de lo que se pensaba originalmente. En esta sección se presentan los primeros pasos en este enfoque de gestión de datos utilizando "*haplotipos*", los cuales incluyen temas de: *variaciones*, *frecuencias* y *poblaciones*. Además, se discute sobre la evolución de la base de datos para apoyar dichos datos. Cada versión nueva del modelo

conceptual introduce cambios en la estructura de la base de datos subyacente, la cual tiene implicaciones esenciales y prácticas para una mejor comprensión y gestión de la información relevante.

Es importante resaltar que facilitar una solución basada en "*modelos conceptuales*" [1] brinda una definición clara del dominio, generando implicaciones directas en el campo médico (por ejemplo, *medicina de precisión*), donde los *Sistemas de Información Genómicos* (GeIS) desempeñan un papel muy importante.

A medida que la aplicación de tecnologías NGS (*Next-Generation Sequencing*) [35] contribuyen a la generación de grandes cantidades de datos (nuevos) como se mencionó anteriormente, para lograr un máximo provecho de todo el conocimiento -*nuevo*- se necesitan construir estructuras que permitan *organizar*, *procesar* y *explotar* los datos, todo esto con el objetivo de mejorar la comprensión del genoma humano. Sin embargo, el modelo conceptual requiere estar constantemente alineado con el nuevo conocimiento genómico, y en este capítulo se extiende el MCGH (v2) antes mencionado para incluir la especificación de los *Haplotipos* (los cuales se definirán en la Sección 5.1), mejorando de esta forma la expresividad del modelo.

Para la incorporación de los "*Haplotipos*" en el MCGH v2, se debe enfocar la ampliación hacia 2 puntos claves:

1. *Integrar el tratamiento de los datos relacionados con los Haplotipos*, y
2. *La aplicación de modelos estadísticos (biológicos)*

De esta manera, prevemos la creación de un *Sistema de Información Genómico* sólido y fiable basado en este modelo conceptual (holístico). Para la consecución de este subobjetivo de la tesis se establecen los siguientes apartados: En la Sección 5.1 se explican los antecedentes asociados al concepto de "*Haplotipo*". La Sección 5.2 comenta los trabajos relacionados en este entorno. En la Sección 5.3 se plantea la propuesta conceptual de extensión del MCGH, posteriormente se presenta la validación del nuevo modelo conceptual (Sección 5.3.1) y el desarrollo de una base de datos de haplotipos (Sección 5.3.2). La Sección 5.4 presenta una comparativa sobre el impacto de la evolución del MCGH sobre la base de datos del genoma humano. Finalmente, la Sección 5.5 presenta las conclusiones del capítulo. Resultados de este capítulo se encuentran publicados en los siguientes trabajos [120], [121] y [122].

5.1 Antecedentes: Comprendiendo el concepto de Haplotipo – caso práctico: Sensibilidad al Alcohol-

En el desarrollo del trabajo investigativo sobre las implicaciones genéticas asociadas a la patología de la "*Sensibilidad al Alcohol*" se detectó la importancia de incluir el tratamiento de Haplotipos en el MCGH. Para ello, se realizó un estudio intensivo de genes y variantes que se asociaron con una predisposición a esta enfermedad [123], [124].

La *sensibilidad al alcohol* se produce cuando un individuo ingiere una cierta cantidad de alcohol, provocando un rechazo inmediato y experimentando molestias, mareos, entre otros síntomas. El simple hecho de consumirlo provoca molestias y tiene un impacto futuro en la salud del individuo [125], [126]. Se ha seleccionado esta enfermedad como caso práctico, debido a que se produce/presenta en la población en su conjunto, independientemente de la condición *social*, *edad* o *cultural* [127].

Durante el estudio de las alteraciones genéticas, se obtuvo una gran cantidad de información provenientes de distintos repositorios de datos genómicos (como, por ejemplo, *NCBI*, mediante sus interfaces: *dbSNP* [71], *PubMed* [128], entre otros). Lo cual permitió definir un grupo de genes (Tabla 6) estrechamente vinculados con la enfermedad [124], entre los cuales se incluyen:

Tabla 6. Genes y variaciones asociadas con la Sensibilidad al Alcohol

GEN	*NOMBRE COMPLETO (OFICIAL)*	*SNP / VARIANTES*
ADHIB	alcohol dehydrogenase 1B (class I), beta polypeptide	rs671; rs1229982; rs1229984; rs1230025
ALDH2	aldehyde dehydrogenase 2 family (mitochondrial)	rs671; rs7590720; rs1800497
GABRA2	gamma-aminobutyric acid (GABA) A receptor, alpha 2	rs279836; rs279858; rs279871
PECR	peroxisomal trans-2-enoyl-CoA reductase	rs7590720
PKNOX2	PBX/knotted 1 homeobox 2	rs1426153; rs750338; rs585977; rs10893366
SLC22A18	solute carrier family 22, member 18	rs16928809
DRD2	DRD2 dopamine receptor D2	rs1076560; rs1800497; rs6276

Estos genes están directa o indirectamente asociados con la *sensibilidad al alcohol*. Los genes con una influencia directa presentan una alta

predisposición con esta enfermedad (ADHIB, ALDH2 y GABRA2), y los indirectos están en una perspectiva más genérica (debido a que están asociados con enfermedades adictivas: PECR, PKNOX2, SLC22A18 y DRD2).

Tras finalizar los procesos de búsqueda e identificación, se procedió a la validación médica de genes, gracias a la ayuda de colegas biólogos del grupo de investigación (los genes validados se soportaron mediante artículos científicos de alto impacto –*revistas médicas*-) [122]. En la Tabla 6 se presenta el filtrado de los genes seleccionados, quedando tres genes relevantes para esta enfermedad: (a) ALDH2 (rs671); (b) GABRA2 (rs279836, rs279871); y (c) ADH1B (rs1229982, rs1229984).

En la realización del estudio se detectó un caso "*haplotípico*" con el gen GABRA2 [129], en el que se encontró un haplotipo compuesto de tres variantes (variaciones): rs279871, rs279836 y rs279845 [120]. Inicialmente se trabajó con variaciones individuales, sin considerar las relaciones entre ellas y sin considerar la variante rs279845 [130], [131].

El rs279871 forma un haplotipo con rs279836 y rs279845

rs279871	RefSNP	Alleles: A/G	Ancestral Allele: A
rs279836	RefSNP	Alleles: A/T	Ancestral Allele: T
rs279845	RefSNP	Alleles: A/T	Ancestral Allele: A

Desde un punto de vista biológico, los *Haplotipos* se definen como un conjunto de SNPs que se heredan y se encuentran juntos en un cromosoma, y son definidos como un grupo de SNPs de un gen que están muy cerca y tienden a ser heredados juntos. Esto significa que los alelos de un haplotipo no se separan en la fase de recombinación y se pueden transmitir en bloques, lo que permite combinaciones de variantes a un gen que afecta a ciertos fenotipos [132].

Actualmente hay un conjunto significativo de enfermedades genéticas en las cuales la influencia de los haplotipos ha sido bien establecida, como, por ejemplo, en el *cáncer de mama* [133], [134] y la *sensibilidad al alcohol* [130], entre otras [135], [136]. Si se consideran los "*haplotipos*" en el diagnóstico, el resultado obtenido podría mejorar, debido a que las probabilidades evaluadas indican el nivel de riesgo en el diagnóstico genómico con mayor precisión.

Tras detectar este nuevo conocimiento, se procedió a estudiar y evaluar el modelo conceptual con el fin de evaluar su estructura y posibles

mejoras, para que el proceso de integración de la información fuera realizado de la manera más apropiada. El correcto tratamiento de este concepto en el modelo era un factor clave para mejorar los resultados generados en los diagnósticos genéticos.

En la Figura 42 se presentan las diferencias encontradas para el análisis genético centrado sólo en variaciones (variantes) individuales frente a los estudios sobre variaciones que participan o forman parte de haplotipos en una muestra de referencia. Como la figura explica (en la ruta A) en esta situación el médico busca un cambio en la muestra asociada con una enfermedad especifica; para este caso simplemente se comprueba si la variación (SNP) existe en la muestra analizada. Con la información obtenida en esta ruta se genera un informe sobre la existencia de la variación en la muestra.

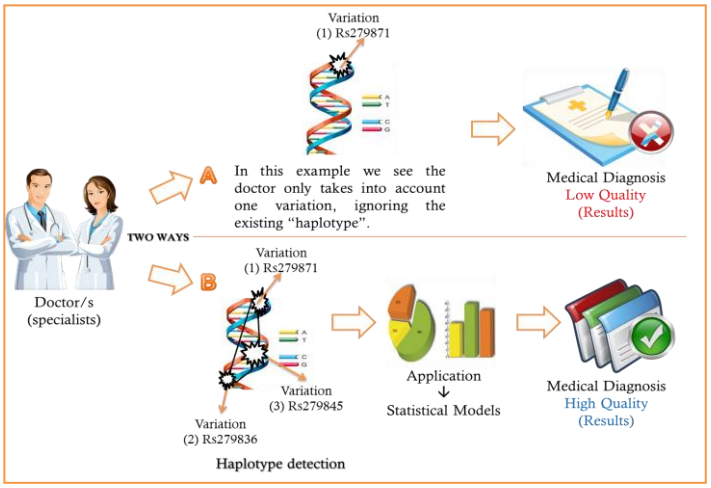


Figura 42. Análisis genético utilizando "*variaciones*" versus "*variaciones + haplotipos*" [120]

La ruta B de la Figura 42, explica el proceso ideal, el cual se presenta en este trabajo. El médico busca las variaciones y determina si hay combinaciones entre todas las variaciones, tratando de encontrar haplotipos en los diferentes alelos mediante el análisis de las frecuencias de cada uno. Para obtener estos datos, es necesario aplicar diferentes modelos estadísticos para presentar un informe genético más detallado y completo. Actualmente es ampliamente aceptado que los estudios de haplotipos mejoran la tasa de detección de variaciones (con o sin combinaciones) para una enfermedad específica [137]. La razón se

debe a que cada alelo o variación representa una frecuencia de ocurrencia en cada población, de modo que con esta información se puede mejorar la generación del diagnóstico genético.

5.2 Trabajos Relacionados

Dentro del estudio realizado al dominio se ha encontrado que poco trabajo se ha hecho hasta la fecha sobre el modelado conceptual de "*haplotipos*". Varios trabajos han tratado de lograr una definición conceptual de todo el genoma humano, pero este conocimiento se vuelve obsoleto rápidamente debido a la continua evolución del dominio.

Esta investigación se centra en la integración y el uso de información existente sobre los haplotipos en los repositorios de información genómica, como, por ejemplo, los sistemas de información y bases de datos. Por estas razones, primero se analizaron algunos de los repositorios genómicos más importantes para evaluar los modelos (esquemas) y los conceptos que utilizan, y cómo se almacenan los haplotipos.

dbSNP: este repositorio es la principal fuente de información sobre SNPs. este repositorio facilita un Esquema E-R (entidad-relación) que identifica la representación de datos sobre: la población, las frecuencias alélicas de un SNP y el resumen de poblaciones. dbSNP recopila dichos datos en la vista "*Frequency calculation submitted by SNP and population*" con la construcción número 118 del 17/11/2003.

Dentro de esta vista se encontró una tabla llamada "*b125_SNPMapInfo_35_1*" relacionada con la tabla "*SNP*", donde sólo hay un atributo relacionado con haplotipos (Tabla 7): "*hap_cnt*" [71], [72], [138].

Tabla 7. Identificador del atributo en dbSNP

Atributo	***Descripción dbSNP***
-hap_cnt	The number of contigs that have the group_term (in *ContigInfo*) with ""*haplotype*"" suffix that the SNP aligns to.

Ensembl: este repositorio proporciona principalmente "*genoma*" para especies de vertebrados. En este repositorio el esquema no proporciona ninguna relación explícita con los haplotipos, pero cabe destacar que en la vista "*features_analyses_core*" se han detectado algunas entidades que podrían estar vinculadas al tratamiento de haplotipos, tales como: la tabla "*Marker_map_location*" y el atributo "*lod_score*", los cuales son datos estadísticos utilizados en la *Genética de Poblaciones* y en los cálculos del LD (de sus siglas en inglés, *Linkage Disequilibrium*) [78], [139].

UCSC Genome Bioinformatics: este sitio contiene secuencias de referencia y conjuntos de borradores (trabajos) realizados, los cuales conforman una gran colección de genomas. En este caso, se presentan los datos como un "*esquema de tablas*", el cual resulta un poco difícil de gestionar [140].

Mediante el uso de la herramienta "*Gene Sorter*" [141], se pudo comprobar los diferentes datos proporcionados para un gen, incluidos los datos asociados con los haplotipos, en la sección "*Common Gene haplotype Alleles*", los cuales son generados a partir del proyecto 1000 Genomas (*http://www.1000genomes.org/*). A partir de los datos estudiados se redactó/diseñó un esquema de la estructura para comparar con la solución planteada en este trabajo [105], [142].

Es importante resaltar que también se encontró un conjunto de bases de datos centradas en la recolección de datos asociados con haplotipos y frecuencias poblacionales, tales como: *HapMap* [83], *ALFRED* [60], *YHRD* [109], *D-HaploDB* [75], entre otras. El problema principal con estos repositorios se basa en la difícil gestión y acceso a la información haplotípica, debido a que dicha información está ampliamente dispersa. Estos datos son almacenados en múltiples archivos de texto (ej. *.txt, *.csv, etcétera). Después de analizar distintas fuentes de datos (y esquemas) que almacenan la información del haplotipo, se identificaron tres problemas principales:

1. **Complejidad en la gestión de datos:**
   Los datos se presentan de forma ambigua, y en muchos casos son difíciles de entender y manipular. En los casos de dbSNP y Ensembl, los datos sobre haplotipos no se muestran explícitamente a los usuarios finales. En este estudio se detectó que el repositorio dbSNP utiliza datos del proyecto HapMap, mientras que USCS utiliza datos del proyecto 1000 Genomas.

También se encontraron algunas incoherencias entre estos repositorios (es decir, información contradictoria o inconsistente en las bases de datos, específicamente entre las frecuencias para alelos y genotipos). Conforme a la evolución continua del entorno genómico, se deben incorporar los conocimientos adicionales en los repositorios de datos. Varias fuentes muestran datos con información sobre haplotipos, como en los casos mencionados anteriormente, pero el problema radica principalmente en la complejidad de la gestión y la interpretación de los datos (ej. *importancia* y/o *relevancia*, etcétera) [143].

2. **Alta dispersión y redundancia de datos:**
   Este problema es una consecuencia de la existencia de diferentes fuentes de datos, las cuales poseen grandes cantidades de información –*estructurada y no estructurada*- presentada en diferentes formatos, como, por ejemplo, **.csv*, **.txt*, **.xml*, **.fasta*, etcétera. Esta amplia gama de formatos hace que sea muy difícil procesar y analizar los datos, por lo que es razonable adoptar en este ámbito los beneficios que ofrecen los "*modelos conceptuales*", ya que permiten crear una estructura en la cual los datos pueden ser compartidos de forma eficaz, y los problemas de redundancia u otras cuestiones pueden reducirse [144].

   Otra desventaja identificada en los datos haplotípicos es su gran dispersión y presencia de datos redundantes [121]. Utilizando un enfoque de modelado conceptual, se busca atender a estos problemas mediante el procesamiento integral de los datos, con el fin de complementar el diagnóstico genético existente actualmente.

3. **No existe una formalización clara del concepto "*haplotipos*":**
   Actualmente, los repositorios analizados no proporcionan una estructura adecuada (modelo/esquema) para la gestión de los haplotipos. En algunos casos, ni siquiera representan el mismo concepto, por ejemplo, se encontraron muchas diferencias entre dbSNP y Ensembl sobre la forma de representar los haplotipos a nivel conceptual (es decir, forma de representación, estructura u otros). Sólo se encontró una especie de especificación en un esquema de tablas por UCSC.

dbSNP sólo presentó un atributo asociado con el concepto de haplotipos en su esquema, y por esta razón es una definición muy limitada, Ensembl, al contrario, no proporciona una definición clara en su esquema. Se encontró el atributo "*lod_score*", el cual se utiliza en genética, pero no específicamente para el tratamiento de los haplotipos. En el caso de UCSC, este repositorio sí que presenta datos sobre haplotipos, pero no presenta un esquema o modelo conceptual.

Existen también otras alternativas para representar el conocimiento en general, y en esta investigación se encontraron "*ontologías*" aplicadas a secuencias biológicas.

Sequence Ontology[20] es un conjunto de términos y relaciones utilizados para describir las características y atributos de secuencias biológicas.

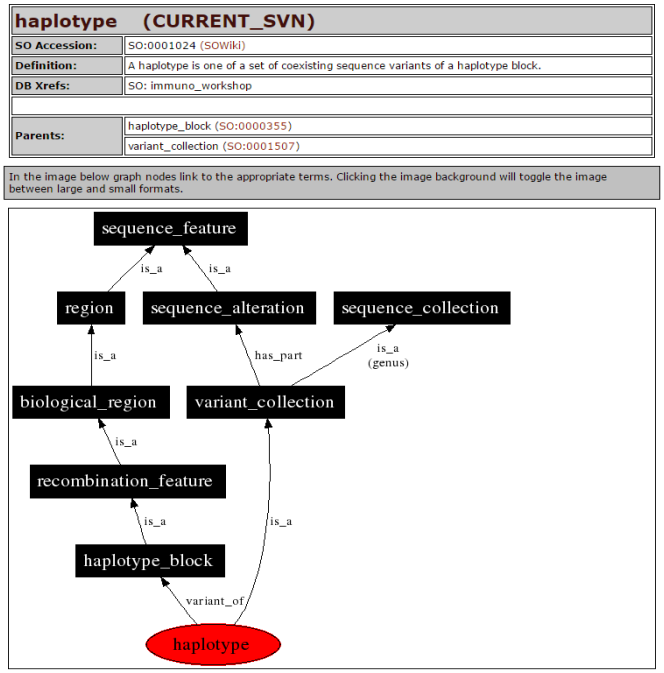


Figura 43. Definición de haplotipos, según *Sequence Ontology* [145]

Esta ontología define un haplotipo como un conjunto de variantes de secuencia coexistente de un bloque de haplotipos (como se puede ver en

[20] *http://www.sequenceontology.org/*

la Figura 43) [145], y este enfoque es interesante para definir los tipos, propiedades y relaciones entre las entidades a un nivel de especificación más formal.

La solución que planteamos busca:

a) Representar los datos existentes en este dominio, manipular y gestionar la información sobre haplotipos para facilitar su uso en el tratamiento genómico (*diagnóstico*),
b) Resolver las deficiencias existentes en este dominio a través de la aplicación práctica de modelos conceptuales, los cuales pueden estar abiertos a la extensión sin importar la evolución continua del dominio genómico.

5.3 Modelado Conceptual de Haplotipos

La aplicación de técnicas de "*gestión de datos*" en un entorno genómico podría verse frustrado o afectado debido a sus características especiales, como son: la alta complejidad a nivel conceptual, las grandes cantidades de datos y la constante evolución del dominio.

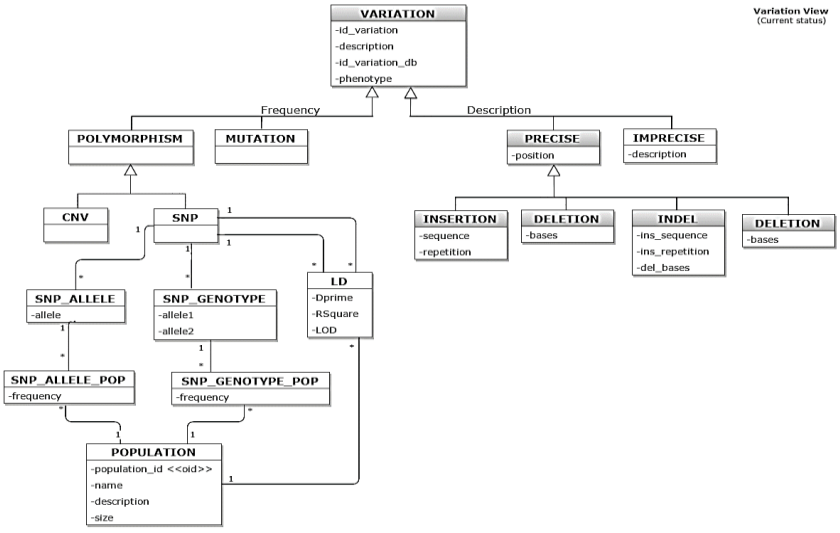


Figura 44. Vista de Variaciones (estado actual) – Fase I

El MCGH descrito en el capítulo anterior ha ido evolucionando a lo largo de los años, y ha sido un gran avance para alcanzar una mejor

comprensión del genoma humano. La idea de integrar los haplotipos al MCGH lo fortalece y consolida, además de evaluar los niveles de incidencia o riesgo de las "*variaciones*" para la predisposición en ciertas enfermedades genéticas. La Figura 44 muestra el estado actual de la vista de variaciones, indicando las clases (*en color gris*) que actualmente están cargadas en el repositorio (utilizando un proceso de carga de datos [146], [147]).

En el MCGH (versión 2) se utilizan las "*variaciones precisas*" (aquellas en que la estructura y los nucleótidos que están involucrados están claramente definidos). Sin embargo, al tratar los haplotipos, otros elementos o conceptos deben ser gestionados, entre los que se incluyen: frecuencias (*alélicas* y *genotípicas*) y poblaciones [84], los cuales son muy difíciles de manejar en el contexto genómico. Este nuevo planteamiento requiere que el proceso de carga de datos también considere los conceptos representados en el lado izquierdo de la Figura 44 (generalización de "*frecuencias*").

En el MCGH (versión 2) las variaciones se presentan en dos grupos, por: *frecuencia* y *descripción*. El primero se clasifica según la frecuencia de la variación dentro de una población (su ocurrencia) y el segundo grupo se ajusta a la descripción dada a las variaciones (pudiendo ser de dos tipos: *precisas* e *imprecisas*), como se ha explicado anteriormente en el Capítulo 4.

La clase "*SNP*" del MCGH se convierte en la clase raíz donde aparecen nuevos conceptos (Figura 45). Aunque la conceptualización de las clases: "*SNP_Allele*"; "*SNP_Allele-Pop*"; "*SNP_Genotype*"; "*SNP_Genotype-Pop*"; "*Population*"; "*LD*" están representadas en la propuesta original, los datos relacionados con las mismas no fueron tratados ni cargados en el repositorio por alguna de estas dos causas:

a) No disponibilidad de los datos (*fuentes*, *recursos*, *etcétera*)
b) No se consideraba que representaran un valor apreciable o tangible

Un SNP se asocia a muchos alelos ("*SNP_Allele*") y estos conjuntos tienen una frecuencia en una población especifica ("*SNP_Allele-Pop*"). Del mismo modo, también posee varios tipos de datos genotípicos ("*SNP_Genotype*") definidos por los atributos "*Allele1*" y "*Allele2*" (el alelo de referencia y el alelo cambiado dentro del genotipo). La clase "*Population*" se utiliza para agrupar todas las poblaciones que se han estudiado para el análisis de variaciones en el genoma humano.

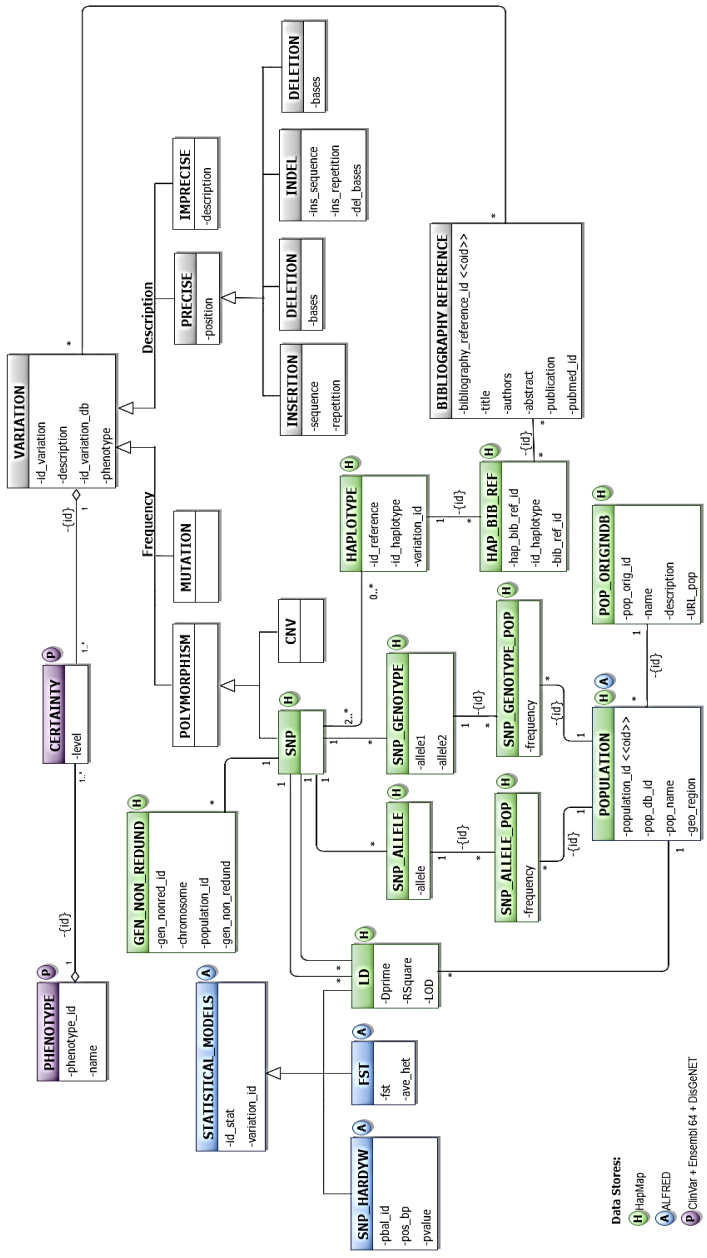


Figura 45. Integración de haplotipos al MCGH – Fase II

La Figura 45 presenta la propuesta desarrollada para integrar los haplotipos en el MCGH, en el cual se encuentran una serie de clases

"*añadidas*" o "*extendidas*". Como se puede observar, la vista de "*Variación*" definida por su "*descripción*" se mantiene como en la versión actual. En esta figura, las variaciones definidas por su "*frecuencia*" tienen que implementar unos cambios necesarios para manejar/gestionar los datos sobre haplotipos e información estadística (modelos estadísticos).

Las inserciones y cambios realizados en el modelo se explican a continuación: se añade la clase "*Haplotype*", la cual está asociada a la clase "*SNP*" representando la combinación de dos o más SNPs. Los atributos que definen la clase "*Haplotype*" en el modelo son: el "*id_reference*" que define un identificador como un enlace entre los diferentes SNPs y haplotipos; los identificadores de los haplotipos y variaciones están representados como "*id_haplotype*" y "*variation_id*" respectivamente.

Como el "*haplotipo*" debe basarse en un recurso científico para corroborar su valor médico y/o científico, se creó una clase llamada "*Hap_Bib_Ref*" como intermediaria y punto de unión entre las clases "*Haplotype*" y "*Bibliography_reference*", logrando de esta forma mantener el repositorio vinculado a varios trabajos de investigación sobre haplotipos.

La clase "*Population*" es reestructurada conforme al nuevo conocimiento a integrar, y esta clase consta de los siguientes atributos: como identificador de la población se utiliza el atributo "*population_id*"; "*pop_db_id*" representa el identificador de la fuente o repositorio de donde se tomó la población; el nombre de la población y de la región geográfica son representados mediante los atributos "*pop_name*" y "*geo_region*" respectivamente.

También se asocia la nueva clase "*Pop_OriginDB*", la cual se utiliza para definir las fuentes que proporcionan las poblaciones antes mencionadas. En esta clase se definen los siguientes atributos: "*pop_orig_id*", el cual es el identificador del repositorio; el nombre ("*name*") y descripción ("*description*") de la fuente, y el "*URL_pop*" que contiene la URL del archivo de datos de la población.

Otra novedad en el modelo es la incorporación de la clase "*Statistical_models*", definida con el objetivo de unificar los modelos estadísticos que se aplican a los datos relacionados con las variaciones [148]–[150], específicamente sobre los cambios en el "*variation_id*",

considerando que para un SNP o variación se pueden tomar de cero a muchos modelos estadísticos. De esta clase salen tres especializaciones, las cuales son: "*SNP_HardyW*", "*Fst*" and "*LD*".

La definición conceptual de estas clases permite abordar conceptos de gran relevancia en el mundo de la genómica, como es el caso de la "*Genética de Poblaciones*". Esta consiste en el estudio de las fuerzas que alteran la composición genética de una especie. Este enfoque se relaciona con mecanismos micro-evolutivos como: *mutación*, *selección natural*, *migración* (flujo genético) y la *deriva genética* [151], [152].

La clase "*SNP_HardyW*" representa el modelo de Hardy-Weinberg (también conocido como *equilibrio panmíctico*). El modelo de *Hardy-Weinberg* se utiliza para calcular las frecuencias genotípicas a partir de las frecuencias alélicas [153], [154], en el que los datos se toman de las fuentes y se aplican al modelo. Esta clase posee las siguientes características: un identificador único de la clase "*pbal_id*"; la posición cromosómica en los pares de bases "*pos_bp*"; y el p-valor ("*pvalue*") que indica el nivel de significación más pequeño posible.

Formula *Hardy-Weinberg*[21]: si consideramos en una población la pareja alélica A1 y A2 de un locus dado,

p es la frecuencia del alelo A1 $0 =< p =< 1$
q es la frecuencia del alelo A2 $0 =< q =< 1$ y $p + q = 1$

Siendo las frecuencias alélicas iguales para ambos sexos, por ejemplo: hombres (p, q) mujeres (p, q)

En la generación siguiente: $(p + q)^2 = \mathbf{p^2 + 2pq + q^2 = 1}$ donde:

p^2 = frecuencia del genotipo A1 A1 ← HOMOCIGOTO
2pq = frecuencia del genotipo A1 A2 ← HETEROZIGOTO
q^2 = frecuencia del genotipo A2 A2 ← HOMOCIGOTO

Estas frecuencias se mantienen constantes de generación en generación.

Otro valor estadístico utilizado en la genética de poblaciones, es el cálculo de los "*Índices de fijación*" [155], los cuales permiten medir la

[21] *http://atlasgeneticsoncology.org/Educ/HardySp.html*

diferenciación de la población por su estructura genética, facilitando la comparación de la variabilidad genética dentro y entre poblaciones.

Para ello se define la clase "*Fst*", compuesta por los atributos "*Fst*" para el índice de fijación y el "*Ave_Het*" que indica la heterozigosidad media. La clase "*LD*" define el "*Linkage disequilibrium*" [156], que se produce cuando los genotipos en los dos loci no son independientes entre sí.

Para obtener el "*LD*", se encontraron tres parámetros biológicos-estadísticos, los cuales son:

1) El índice de sensibilidad "*DPrime*", mide el desequilibrio entre la interacción de los alelos.
2) El coeficiente de determinación "*RSquare*" [157], sirve para determinar la calidad del modelo, con el objetivo de replicar los resultados, y la proporción de variación en los resultados presentados en el modelo.
   Para su cálculo, por ejemplo, se muestran dos casos: (1) caso general y (2) para la regresión lineal.

   (1) Si se representa por σ^2 la varianza de la variable dependiente y la varianza residual por σ_r^2, el coeficiente de determinación viene dado por la siguiente ecuación:

   $$\rho^2 = 1 - \frac{\sigma_r^2}{\sigma^2}$$

   Esta se mide en tantos por ciento. Si la varianza residual es cero, el modelo explica el 100% de valor de la variable; si coincide con la varianza de la variable dependiente, el modelo no explica nada y el coeficiente de determinación es del 0%.

   (2) Para la regresión basta con hacer el cuadrado del coeficiente de correlación de Pearson.

   $$R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2}$$

   Donde:
   σXY es la covarianza de (X, Y)
   σX es la desviación típica de la variable X
   σYes la desviación típica de la variable Y

3) LOD Score "*LOD*" [158], este valor se refiere al logaritmo en la probabilidad de que dos genes o loci estén enlazados y por lo tanto se hereden juntos más a menudo que de costumbre.

La clase "*LD*" está relacionada con la clase "*SNP*", la cual indica que un SNP puede tener de cero a muchos LDs. En esta extensión del MCGH se integran (carga) todos los datos existentes sobre "*Fenotipos*", los cuales proceden de diferentes repositorios, para ello se toma la clase "*Phenotype*" y se asocia a la clase "*Variation*" mediante la clase intermedia "*Certainty*", esta última utilizada como indicador del nivel de incidencia entre *fenotipo-variación* (este valor es extremadamente difícil de definir, pero los estudios sobre este tema proporcionan "*valores estimados*" dentro de la población estudiada).

Por último, se incorpora la clase "*Gen_non_redund*", la cual ayuda a ofrecer los resultados de los conjuntos de datos curados (sin redundancia, es decir, eliminación de inconsistencias y datos duplicados –*datos extraídos de HapMap*-), para conjuntos de SNP-genotipificación y población. Para esta se clase se asignó un identificador único para los datos –*sin redundancia*- "*gen_nonred_id*"; información sobre el cromosoma y la población estudiada, para ello se utilizan los atributos "*chromosome*" y "*population_id*" respectivamente; y para el total de los datos no redundantes (después de realizar el filtrado) se utiliza el atributo "*gen_non_redund*".

5.3.1 Validación del Modelo Conceptual

Para validar este trabajo, se verificó que el modelo conceptual -*extendido*- soporte la información proporcionada por los repositorios típicos sobre haplotipos, estableciendo una alineación conceptual con los datos disponibles en los repositorios genómicos –*más relevantes*-.

La Figura 45 presenta los elementos añadidos al modelo y sus respectivas fuentes. Cabe destacar que la principal fuente de datos utilizada fue la facilitada por el *Proyecto HapMap* en su tercera fase [159]. De acuerdo con la vista de variaciones, el conjunto de datos provenientes de HapMap están contenidos en los siguientes directorios:

a) *Frequencies*, y
b) *LD_data*

En primer lugar, "*Frequencies*" es un directorio que contiene los SNP anteriores con su frecuencia para cada población. Debido a la gran cantidad de información, se agrupa en dos grupos: (1) las frecuencias de las variaciones (SNP), teniendo en cuenta sólo un alelo del cromosoma ("*Allele_freqs*") y las frecuencias de las variaciones, teniendo en cuenta los dos alelos ("*Genotype_freqs*").

En la Tabla 8 se muestran los elementos añadidos al modelo, los repositorios de datos y el campo/tabla específica que está alineada con el nuevo elemento.

Tabla 8. Elementos del modelo + Fuentes de datos (origen)

	Schema element (name)	Origin	
		Data source	File / Table
"*frequencies*" and "*LD_data*"	SNP_Allele	HapMap	*Allele_Freqs_X_Y*
	SNP_Allele-pop	HapMap	*Allele_Freqs_X_Y* → X: *Chromosome* Y: *Population*
	SNP_Genotype	HapMap	*Genotype_Freqs_X_Y*
	SNP_Genotype-pop	HapMap (*BioQ*)	*Overall Hardy-Weinberg*
	Gen_non_redund	HapMap	*Genotyped non-redundant QC+ SNPs*
	Population	HapMap; ALFRED	*List of populations*
"*frequencies*"	Statistical_models	ALFRED	*Siteswithfstavghet*
	Fst	ALFRED	…
	SNP_HardyW	ALFRED	*Overall Hardy-Weinberg*
	LD	HapMap	*LD_X_Y*
	Phenotype	ClinVar; Ensembl; DisGeNET	*ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/ Ensembl 64 phenotypes all_gene_disease_associations.tsv*

Los datos sobre el "*Linkage Disequilibrium*" [160] es proporcionado mediante el directorio "*LD_data*". La gran cantidad de información está dividida por cromosomas y poblaciones.

La información sobre poblaciones "*Population*" y los genes no redundantes "*Gen_non_redund*" se encuentran en secciones específicas de la página principal (*proyecto HapMap*) *-la* Figura 46 *presenta el modelo Entidad-Relación inicial-*.

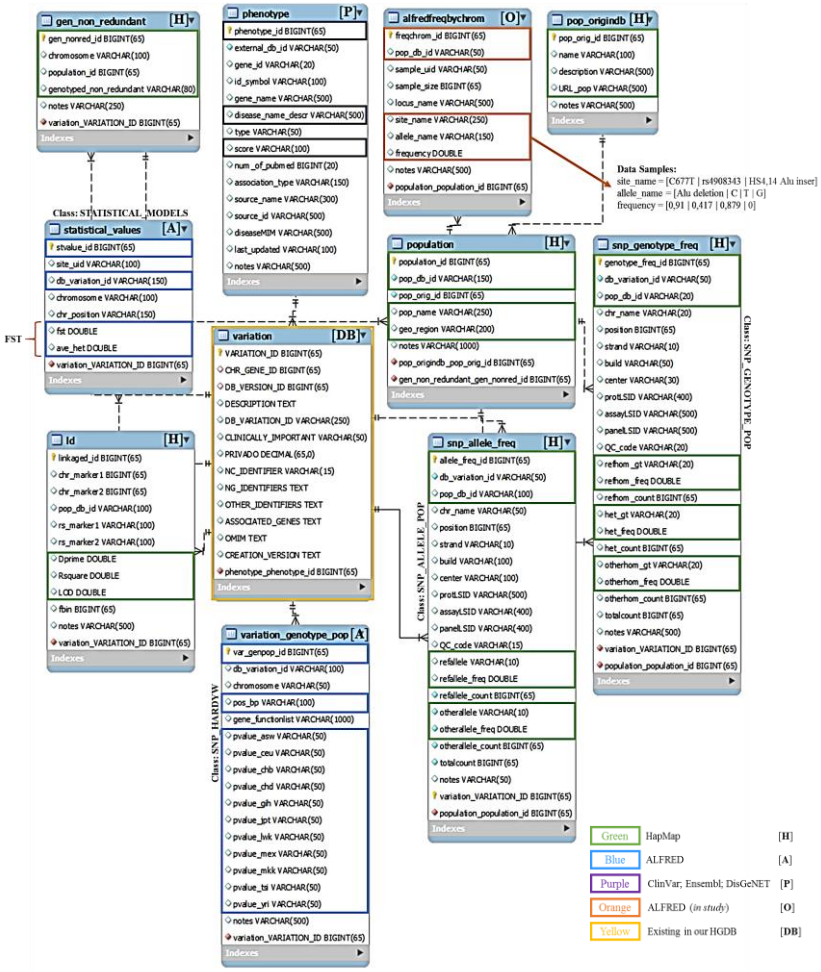


Figura 46. Modelo Entidad-Relación (inicial)

La clase "*Statistical_models*" consiste en los datos extraídos de HapMap para el "*LD*", y el repositorio ALFRED para las frecuencias y procesos estadísticos aplicados al campo biológico. El dataset de frecuencias de ALFRED se obtiene de un archivo llamado "*FreqByChrom*" disponible en su sitio web (*https://alfred.med.yale.edu/alfred/alfredDataDownload.asp*); estas frecuencias se han obtenido para "*cromosomas*" según la población estudiada. Este repositorio además facilita la extracción de la información relacionada con las poblaciones que han utilizado.

Para la entidad "*fenotípica*", se detectaron varias fuentes de datos a partir de las cuales se pueden extraer toda la información, esta se obtuvo de múltiples repositorios, tales como: ClinVar [65], Ensembl [78] y DisGeNET [161], [162] (*ordenados según su relevancia científica*).

5.3.2 Desarrollo de una Base de Datos de Haplotipos

En esta sección se presenta el tratamiento de los datos (*haplotipos*) del modelo conceptual previamente definido (Figura 45), los datos fueron extraídos de los distintos repositorios de datos mencionados en la sección anterior, con el objetivo de definir un conjunto de consultas para el análisis de los datos.

La propuesta planteada para la gestión de los datos se compone de los siguientes pasos:

a) *Estudio de impacto de los repositorios existentes/disponibles*:
Esta actividad se centra en evaluar la utilidad de los datos obtenidos, y su impacto en la investigación (y en las publicaciones científicas), con el objetivo de definir una estructura (*framework*) de fuente de datos acorde con los últimos avances en el contexto genómico.

b) *Recolección de datos (frecuencias y cálculos estadísticos para haplotipos)*:
Después de seleccionar los repositorios de datos, se procedió a descargar los archivos (especificados en la Tabla 8). En esta etapa es habitual encontrarse con grandes archivos de datos, como, por ejemplo:
   - Para las frecuencias de los alelos en el cromosoma 1 con las poblaciones estudiadas, en este caso de HapMap[22] (*11 poblaciones*) se encontraron ficheros con un promedio de 450-1,024MB ±.
   - En el caso de los datos del "*LD*" para el cromosoma 1 y la población "*ASW*" (*African ancestry in Southwest USA*) obtuvo un promedio de 930-1,024MB ±.
   - En el caso del repositorio ALFRED, se obtuvieron las frecuencias cromosómicas con un promedio de 500-

[22] *http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html*

650MB ±. Los archivos con datos fenotípicos eran menores y bastante asequibles en cuestiones de tamaño.

Tras finalizar el proceso de extracción de todos los ficheros, se generaron múltiples gigas de información dispersa y con estructuras completamente distintas para ser analizadas/evaluadas.

Con el fin de reforzar los análisis previamente realizados, se decidió complementar y comparar la información con los datos generados en la plataforma *BioQ* (*http://bioq.saclab.net/*) [64], la cual proporciona un conjunto de herramientas para *consultar*, *documentar*, y *descargar* información de bases de datos relacionales (genómicas), tales como: *1000 genomas*, *dbSNP*, *Ensembl*, entre otros (*para versiones específicas que ellos manejan*).

c) *Análisis de los datos almacenados*:
Debido a la heterogeneidad de los datos, lo primero que se hizo en esta fase fue "*transformar*" los datos de su formato actual (ver Sección 5.2) a una estructura común. En este paso se decidió convertir todos los ficheros en formato "*.csv" para su depuración y análisis, y luego se tomaron los cromosomas (del 1 al 3) del genoma como caso de prueba, esto por cuestiones de manipulación y almacenamiento (grandes gigas de datos).

Como se puede observar en las Figuras 47 y 48, después de finalizar el análisis de los datos (relevantes) para el tratamiento de los haplotipos y modelos estadísticos, se identificó la gran contribución de los repositorios, como, por ejemplo, *ALFRED* (1.063.651 filas) y *HapMap* (194.417 filas) para este estudio.

Además, se clasificó todo el conocimiento obtenido de acuerdo con cinco categorías: *frecuencias*, *valores estadísticos*, *fenotipos*, *poblaciones* y *otros*. La mayor cantidad de datos procesados se encontraron en frecuencias y valores estadísticos para las variaciones.

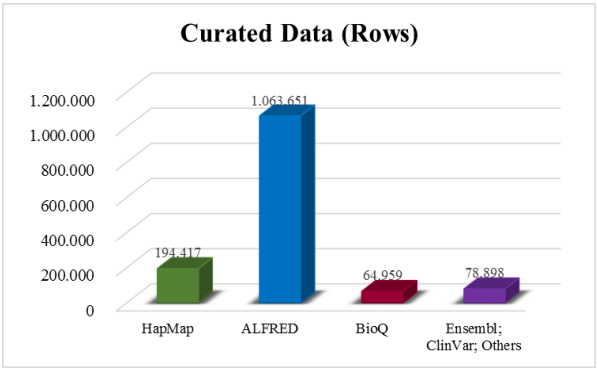


Figura 47. Datos curados cargados en el repositorio de datos

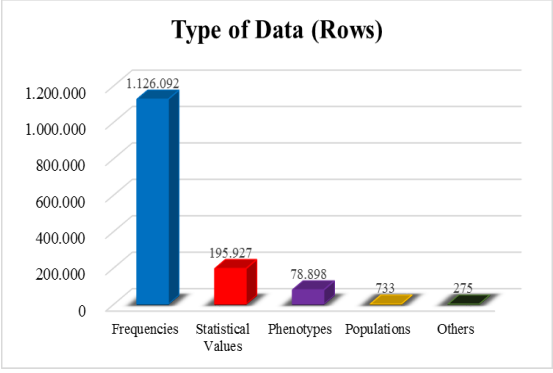


Figura 48. Tipos de datos almacenados (total de filas)

d) *Carga de datos masiva*:
Tras finalizar los procesos de "*análisis*" y "*tratamiento*" de los datos, se realizó un filtrado preliminar de los mismos, y como siguiente paso se desarrolló un esquema de base de datos inicial para proceder a cargar toda la información. Para esto, se estudiaron las estructuras de datos más apropiadas para organizar, consultar y evitar problemas de procesamiento (por ejemplo, *signos especiales*) en los datos.

La siguiente tarea consistió en importar los archivos previamente generados (*.csv) utilizando un entorno de administración de bases de datos, en este caso *HeidiSQL*[23] (ver Figura 49). Con esta tarea completa se generó un nuevo

[23] *https://www.heidisql.com/*

diagrama E-R manteniendo la trazabilidad del origen de los datos cargados (Figura 46).

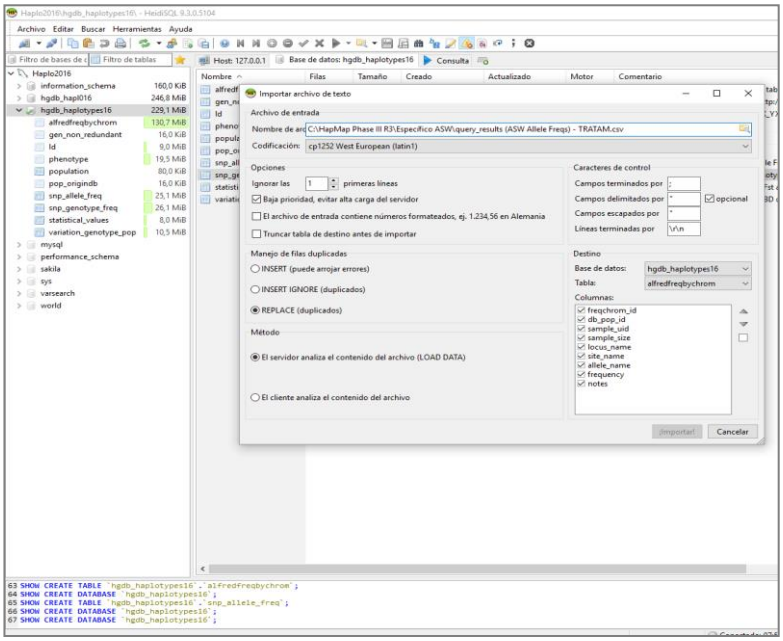


Figura 49. Importación de datos utilizando HeidiSQL

e) *Generación de consultas (preliminares)*
En esta fase se plantearon una serie de preguntas con el objetivo de encontrar respuestas mediante la generación de diferentes consultas SQL en el esquema previamente cargado, dentro de las cuales tenemos como ejemplos:

```
(1) SELECT COUNT(DISTINCT(id_symbol)) FROM phenotype
```

Con esta consulta (1) se encontró un total de 4.192 genes cargados, los cuales representan una asociación con una o más enfermedades o fenotipos.

```
(2) SELECT phenotype_id, id_symbol, disease_name_descr, type, score, num_of_pubmed, source_name, diseaseMIM, last_updated, notes FROM phenotype WHERE id_symbol = 'GABRA2'
```

phenotype_id	id_symbol	disease_name_descr	type	score	num_of_pubmed	source_name	diseaseMIM	last_updated	notes
72.387	GABRA2	Alcohol dependence	disease		0	NCBI curation	103780	26-oct-11	ClinVar

Con esta consulta (2) se obtuvo el gen "*GABRA2*" con el identificador 72.387 ("*phenotype_id*"), el cual fue cargado desde la fuente de datos de "*ClinVar*" y está asociado con la "*dependencia del alcohol*" (ver Sección 5.1).

Otras consultas relevantes fueron sobre los temas de: *frecuencias de variaciones*, *poblaciones*, y *representación de datos estadísticos* (biológicos) obtenidos para el estudio.

```
(3) SELECT allele_freq_id, db_variation_id, pop_db_id, chr_name, position, strand, refallele, refallele_freq, otherallele, otherallele_freq, totalcount FROM snp_allele_freq WHERE position BETWEEN 9200000 AND 9209000
```

allele_freq...	db_variatio...	pop_db_id	chr_na...	position	strand	refallele	refallele_f...	otherallele	otherallele_f...	totalcount
4.831	1009940	ASW	ch1	9.200.729	+	G	0,711	A	0,289	114
4.832	17033526	ASW	ch1	9.202.760	+	A	0,875	G	0,125	112
4.833	7534423	ASW	ch1	9.203.235	+	G	0,877	T	0,123	114
4.834	12402600	ASW	ch1	9.204.303	+	G	0,763	A	0,237	114
4.835	10489436	ASW	ch1	9.208.308	+	A	0,702	G	0,298	114
4.836	12072683	ASW	ch1	9.208.605	+	C	0,491	T	0,509	114

Para reducir la cantidad de datos en esta consulta, se estableció un rango de búsqueda. Aquí se presentan seis variaciones ("*db_variation_id*") incluidas en la posición del cromosoma 1, comprendida entre -9.200.000 y 9.209.000- (en la cadena positiva), para la población asiática (ASW).

Para este caso específico de la variación "*1009940*", se tiene como alelo de referencia "*G*", el cual tiene una frecuencia de "*0,711*", y el otro alelo (cambiado/alterado) "*A*", posee una frecuencia de "*0,289*" en esta población (con un conteo total de 114 casos) y así sucesivamente para cada variación.

En el estudio de los diferentes datos estadísticos, para los cálculos relacionados con los datos del *LD* se incluyen variaciones y poblaciones. La próxima consulta (4) presenta una selección de los datos más frecuentes y repetidos en la tabla "*LD*", ordenados según el número de repeticiones (columna "*num*"):

```
(4) SELECT pop_db_id, rs_marker1, rs_marker2,
    Dprime, Rsquare, LOD, fbin, COUNT(*) AS num
    FROM ld GROUP BY 'rs_marker1' ORDER BY num DESC
    LIMIT 0 , 15
```

pop_db_id	rs_mark...	rs_mark...	Dprime	Rsquare	LOD	fbin	num
CHD	16919558	2804311	0,078	0	0	5	251
CHD	2804311	2641984	1	0,252	7,11	5	250
CHD	2641984	2641983	1	0,261	7,21	5	249
CHD	2641983	7031553	1	1	30,2	5	248
CHD	7031553	9632892	1	0,128	3,4	5	247
CHD	7048037	10975061	1	0,937	26,96	5	247
CHD	2641989	16919558	1	0,005	0,31	5	247
CHD	10975061	7040388	1	0,906	25,24	5	246
CHD	1565793	10815231	1	0,159	4,11	5	246
CHD	9632892	1565793	1	1	34,56	5	246
CHD	9408625	7048037	1	0,968	28,4	5	245
CHD	10815231	10975130	1	0,97	31,24	5	245
CHD	10758683	9408630	1	0,318	8	5	245
CHD	7040388	10758683	1	0,288	6,92	5	245
CHD	2804313	2279619	1	0,332	10,05	5	245

Como se mencionó en la Sección 5.3, el "*Linkage disequilibrium*" (LD) permite identificar cuando los genotipos en los dos loci[24] no son independientes entre sí, y se basa en modelos estadísticos utilizados en la *genética de poblaciones*, como: *DPrime*, *Rsquare* y *LOD* (descritos en la Sección 5.3).

```
(5) SELECT linkaged_id, pop_db_id, rs_marker1,
    rs_marker2, Dprime, Rsquare, LOD, fbin FROM ld
    WHERE rs_marker1 = 16919558
```

linkaged_id	pop_db_id	rs_marker1	rs_marker2	Dprime	Rsquare	LOD	fbin
57.936	CHD	16919558	2804311	0,078	0	0	5
57.937	CHD	16919558	2641984	1	0,014	0,38	5
57.938	CHD	16919558	2641983	0,057	0	0	5
57.939	CHD	16919558	7031553	0,068	0	0	5
...	...	...	...	...	...	...	...

[24] Un locus es el lugar específico del cromosoma donde está localizado un gen u otra secuencia de ADN, como su dirección genética. El plural de locus es "*loci*" [199].

Esta consulta indica, por ejemplo, que en individuos de la población "*CHD*" (Beijing, China) tomando las variaciones '*16919558*' y '*2804311*', encontramos un DPrime = 0.078, RSquare y LOD = 0. Estos datos permiten decir que estas variaciones aplicadas al modelo estadístico no son muy dependientes entre sí, lo que indica igualmente una probabilidad de que existan otras variaciones con mayor complicidad.

Además, al lanzar la consulta (5) se pudo observar el cálculo del LD para la variación "*16919558*" ("*rs_marker1*"), donde se compara con un total de 251 variaciones ("*rs_marker2*") para la población "*CHD*" (*Chinese in Metropolitan Denver, Colorado*).

En el tratamiento y estudio de estos datos se ha llevado a cabo cada paso con diferentes parámetros de control con el fin de generar un resultado fiable y sólido.

Es importante destacar de que, a pesar de la gran heterogeneidad y dispersión de los datos existentes en los distintos repositorios analizados, se pueden reducir estos problemas (datos en bruto) mediante la recopilación del conocimiento, la aplicación de técnicas de modelado conceptual y gestión de datos, para lograr generar y gestionar repositorios con datos curados.

En este trabajo de investigación se buscó explotar estos datos sobre haplotipos de una manera distinta y nueva, para sacar provecho a los conocimientos actuales sobre *variaciones genéticas* y temas relacionados con la *genética de poblaciones*. Este conocimiento colabora positivamente en la detección de enfermedades genéticas, con especial énfasis en la "*Medicina Personalizada*" (o medicina de precisión).

5.4 Evolución de la Base de Datos según el Modelo Conceptual

Este apartado se centra en explicar la forma en que influyen o afectan las diferentes representaciones del conocimiento genómico en las estructuras de las bases de datos utilizadas para manejar los datos. Esta es una parte importante en este trabajo, como se ha explicado anteriormente, los datos están disponibles en diferentes fuentes de datos genómicas y la estructura de base de datos seleccionada determina como se van a gestionar los datos.

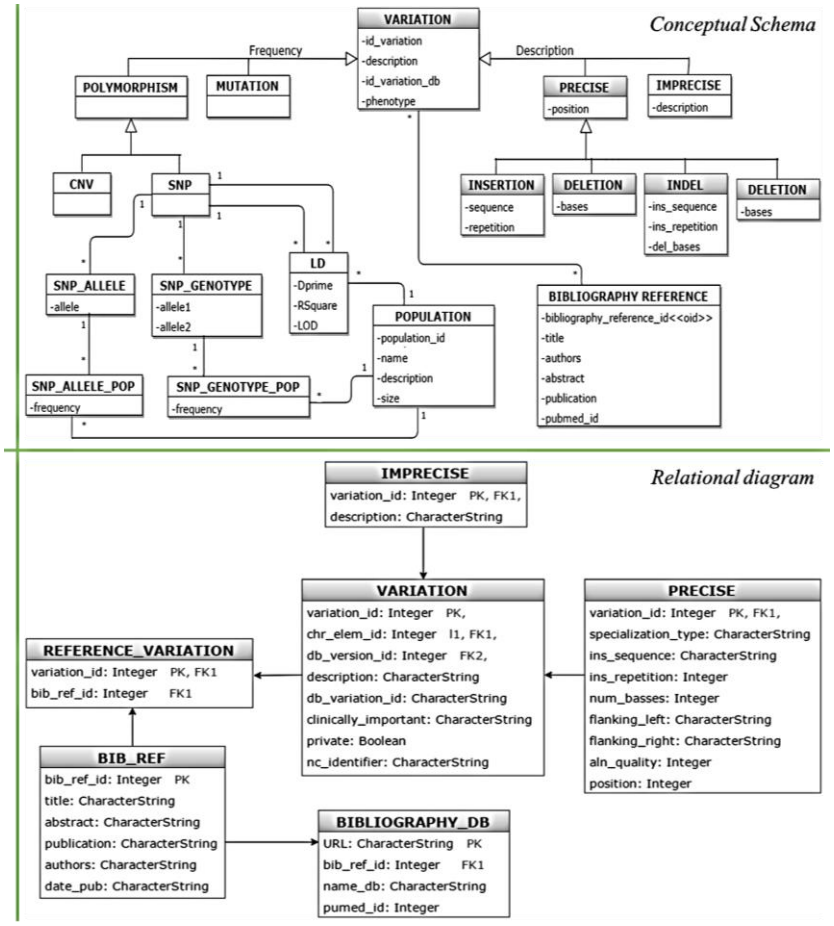


Figura 50. Versión anterior (actual)

El papel importante en este contexto es mantener una perspectiva conceptual –*global*- del problema (*gestión de datos*), independientemente de la estructura de la base de datos, para de esta forma obtener un banco de trabajo conceptual preciso que permita integrar los datos correctamente.

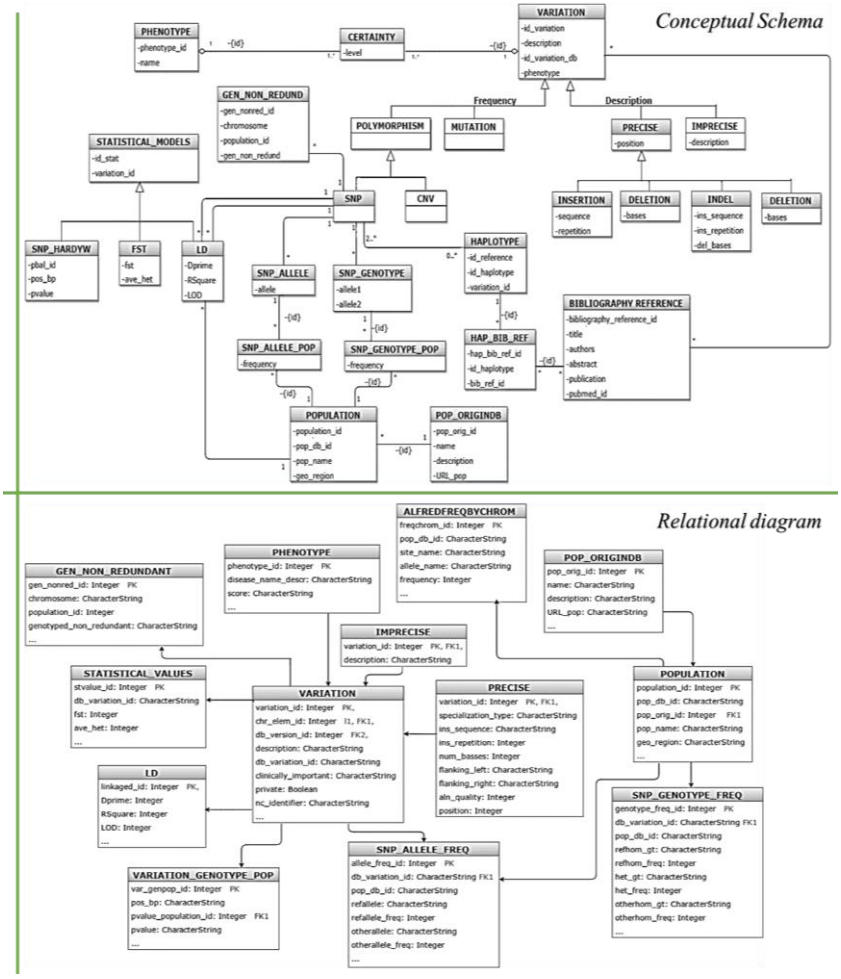


Figura 51. Nueva versión (*extensión*)

Las Figuras 50 y 51 presentan un fragmento de un diagrama de base de datos relacional generado a partir de la versión existente (anterior, versión 2) y la propuesta planteada (versión extendida) en el MCGH que representa los conceptos utilizados en esta vista.

La versión anterior sólo representa los datos cargados en el repositorio genómico, variaciones definidas por la especialización "*Description*", este diagrama relacional se compone de las definiciones "*precisas*" e "*imprecisas*" (con sus respectivas referencias), aquí se omite la parte relacionada con *frecuencias* y *poblaciones* (no almacenadas). Como se ha explicado en este capítulo, los haplotipos deben integrarse en el modelo conceptual porque mejoran en gran medida la expresividad y detalle de los diagnósticos genéticos.

La Figura 51 muestra el modelo conceptual extendido con la integración de haplotipos. En este caso, se incorporó todos los elementos de la variación, por "*descripción*" y "*frecuencia*", generando en este orden un diagrama relacional más completo. Esta nueva forma de representación busca reducir los problemas de dispersión y prevenir la redundancia.

Esta representación trata las variaciones de manera más específica debido a que cubre factores importantes en el conocimiento actual del dominio. Cuando el diagrama identifica las relaciones entre las variaciones como parte de un "*haplotipo*" en el genoma, la información es mejorada por incorporar un mayor nivel de detalle, y además apunta a resultados más precisos.

Los datos poblacionales son interesantes para los diagnósticos, por lo que concretar relaciones entre variaciones y poblaciones juega un papel fundamental por su gran impacto en los resultados que obtendrán los usuarios finales (medicina personalizada). Esta forma de representar los datos relacionados a las variaciones del genoma humano define un nuevo modelo "*Conceptual*" y "*Lógico*", el cual es muy útil para gestionar la información (carga/manipulación), además de proporcionar una mejora sustancial en el rendimiento.

Es importante resaltar que cuando se facilita una representación exacta de los datos involucrados en el genoma humano, se obtiene un modelo conceptual preciso destinado a guiar las diferentes posibilidades potenciales para el almacenamiento y gestión de los datos. En otras palabras, aunque se puedan plantear diferentes representaciones de las mismas fuentes de datos, sólo con un modelo conceptual bien definido se puede proporcionar una perspectiva conceptual unificada, la cual es necesaria para manipular correctamente los datos y comprender la forma en que son almacenados.

5.5 Conclusiones

En el presente capítulo se ha explicado el - *¿por qué*? - de la necesidad de extender el *Modelo Conceptual del Genoma Humano* (MCGH) con la integración de los haplotipos, resaltando que esta representación conceptual del genoma es extensible y permite la inclusión de nuevo conocimiento conservando su definición inicial.

En primer lugar, se explicó el significado biológico del concepto de "*Haplotipo*[25]", y su relevancia dentro del dominio genómico. Este nuevo conocimiento surgió a partir de la aplicación de la Metodología SILE en el caso práctico de la "*Sensibilidad al Alcohol*", pasando posteriormente a realizar la búsqueda de otros casos haplotípicos en genes como *BRCA1* y *BRCA2* asociados con el *cáncer de mama*. Todo esto con el objetivo de conocer su repercusión médica y justificar su integración en el modelo.

Para ello, fue necesario estudiar y analizar distintos enfoques de representación para detectar si el concepto de haplotipos era considerado en los repositorios genómicos ampliamente utilizandos (por ejemplo, dbSNP, Ensembl y UCSC). Además, de evaluar el enfoque ontológico facilitado por la plataforma *Sequence Ontology*. En este punto se encontró que para los haplotipos salían a la luz los típicos problemas del caos de datos genómicos: a) *complejidad en la gestión de los datos*; b) *alta dispersión y redundancia*; y c) *que no existe una formalización clara del concepto*.

En este capítulo se planteó la extensión del MCGH, incorporando este nuevo conocimiento (y todos sus conceptos asociados) dentro del modelo, como, por ejemplo, temas de *frecuencias alélicas*, *datos poblacionales*, entre otros. Se realizaron pruebas de validación con los datos obtenidos y su aplicación en el modelo, así como el desarrollo de una base de datos de haplotipos. Esta última con el objetivo de analizar los datos y ver su aporte en la generación de diagnósticos genómicos.

Finalmente, se presentó una comparativa de cómo evoluciona la base de datos del genoma humano (HGDB) de acuerdo con el crecimiento

[25] Los *Haplotipos* se definen como un conjunto de SNPs que se heredan y se encuentran juntos en un cromosoma, y son definidos como un grupo de SNPs de un gen que están muy cerca y tienden a ser heredados juntos.

del MCGH. Todo esto confirma que esta representación conceptual permite *integrar*, *gestionar* y *entender* el conocimiento genómico existente.

El próximo capítulo tiene como fin presentar a nivel práctico la implementación de un prototipo basado en el MCGH definido, el cual permite gestionar los datos obtenidos a través de un conjunto de repositorios genómicos. Mediante esta solución se contextualiza el último subobjetivo planteado en esta Tesis Doctoral.

CAPÍTULO 6

Implementación

El objetivo de este capítulo es presentar el mecanismo de evaluación implementado, el cual está basado en el MCGH descrito en el Capítulo 4. Como instrumento de implementación del trabajo se ha desarrollado un *prototipo* enfocado en una vista del modelo conceptual "*Variaciones*", donde se demuestra que por medio de la aplicación de modelos conceptuales se pueden desarrollar *Sistemas de Información Genómicos* que permitan manipular las grandes cantidades de información genómica existente alrededor del mundo.

Cuando se emplean técnicas de *Ingeniería de Software* (IS) en el dominio genómico se contribuye a la agilización de la explotación del conocimiento genómico que surge día a día, lo cual brinda mejores mecanismos de distribución y acceso a la tan aclamada "*Medicina de Precisión*".

En primer lugar, en la Sección 6.1, se introduce la metodología SILE aplicada para el estudio de las enfermedades de origen genético. La Sección 6.2 presenta la base de datos del genoma humano (HGDB, *Human Genome Database*) que se generó a partir del MCGH. En la

Sección 6.3 se presenta el desarrollo de la solución del proyecto, llamada "*VarSearch*". Por último, en la Sección 6.4 se introduce un proyecto paralelo desarrollado para ofertar *Test Genéticos Directos al Consumidor* (TGDC), permitiendo de esta manera explotar el nuevo conocimiento a través *Diagnósticos Genómicos* que garantizan la aplicación de la *Medicina de Precisión* (PM). Resultados de este capítulo se encuentran publicados en los siguientes trabajos [122], [163], [164] y [165].

6.1 Metodología SILE

El dominio bioinformático exige tener un control estricto en la manipulación de los datos, debido a que sus principales investigaciones y resultados están basados en cierta medida al conocimiento ya existente y localizado en las grandes cantidades de información (datos) facilitados por los distintos repositorios de datos genómicos. Por lo que es necesario realizar un estudio a fondo que permita garantizar resultados fiables.

Los repositorios de datos genómicos (ej. *NCBI*, *OMIM*, etc.) proporcionan un conjunto extenso de información, y es importante poder extraer información concisa para contribuir en la generación de resultados precisos.

El objetivo de la *Metodología SILE* (de sus siglas en inglés, "*Search-Identification-Load-Exploitation*") es facilitar una plantilla bioinformática para la obtención y tratamiento de los datos provenientes de los repositorios de datos existentes, contribuyendo así en la mejora de los procesos de "*Búsqueda-Identificación-Carga-Explotación*" de la información genómica.

Esta metodología fue planteada por el Prof. Dr. Óscar Pastor López, director del Centro de Investigación PROS de la Universitat Politècnica de València, con el fin de mejorar los procesos de carga de genes y variaciones en la Base de Datos del Genoma Humano (HGDB) [122].

Sus iniciales se definen como sigue a continuación:

- **S**: Búsqueda ("*Search*")
  Consiste en la búsqueda exhaustiva de información científica (*publicaciones*, *artículos*, *etcétera*) para apoyar la asociación genética con una enfermedad especifica.

- **I**: Identificación ("*Identification*")
  Es el proceso donde entra la intervención médica (*experta*) para proporcionar apoyo en el filtrado de los genes y variaciones que poseen mayor impacto (*incidencia*) dentro de la población. Cabe destacar que en los trabajos realizados dentro del *Grupo Genoma del PROS* siempre se ha contado con la colaboración de profesionales de las áreas de biología molecular, biomedicina y biotecnología para la validación final de los datos (*variaciones, genes, etcétera*).

- **L**: Carga ("*Load*")
  Esta tarea consiste en el proceso de carga de la base de datos (HGDB), aquí mediante un proceso ETL (que incluye diversas fuentes de datos) se procede a insertar los genes, cromosomas y variaciones en la base de datos con la información que ya ha sido tratada y validada. La carga se completa tras finalizar distintas tareas.

  Es importante resaltar que para los procesos de carga podemos distinguir dos tipos: (a) *Masiva*, y (b) *Selectiva*.
  - La carga *masiva* consiste en la extracción de todo el conocimiento existente en los distintos repositorios genómicos, donde luego los datos son cargados en la base de datos sin ningún tratamiento o estudio previo (son los datos en bruto).
  - La carga *selectiva* consiste en la búsqueda de información específica en los distintos repositorios genómicos, dicha información está relacionada con variaciones que afectan una enfermedad predeterminada (procesos "*Search-Identification*" de la metodología SILE). Este método de carga permite generar extraer la información deseada y generar bases de datos curadas (*limpias*, *fiables*, *etcétera*). El objetivo es mantener datos relevantes y que añadan valor

(mediante referencias, soporte científico) al diagnóstico genómico (resultado).

- **E**: Explotación ("*Exploitation*")
  Esta tarea se basa en la representación del contenido, es la presentación del resultado final.

Figura 52. Metodología SILE (*Search-Identification-Load-Exploitation*)

La metodología SILE (como se puede ver en la Figura 52) se aplica a bases de datos genómicas que facilitan la búsqueda y consulta de información biomédica (como, por ejemplo, las presentadas en la Sección 6.1.1).

Dentro de los trabajos desarrollados utilizando la metodología SILE se pueden encontrar una amplia variedad de enfermedades estudiadas, como, por ejemplo, *sensibilidad al alcohol* [122] [127], *dupuytren* [166], *alopecia androgénica* [167], *intolerancia a la lactosa* [168], etcétera. Hoy en día se continúa aplicando la metodología SILE a distintas enfermedades como: *neuroblastoma* [165], *alzhéimer*, *cataratas*, *cáncer de pulmón*, entre otras.

6.1.1 Ejemplos de búsquedas en repositorios genómicos

A continuación, se presentan algunas de las pantallas más reconocidas y utilizadas por los genetistas y expertos al momento de buscar información sobre *cromosomas*, *genes*, *variaciones*, *fenotipos*, entre otros.

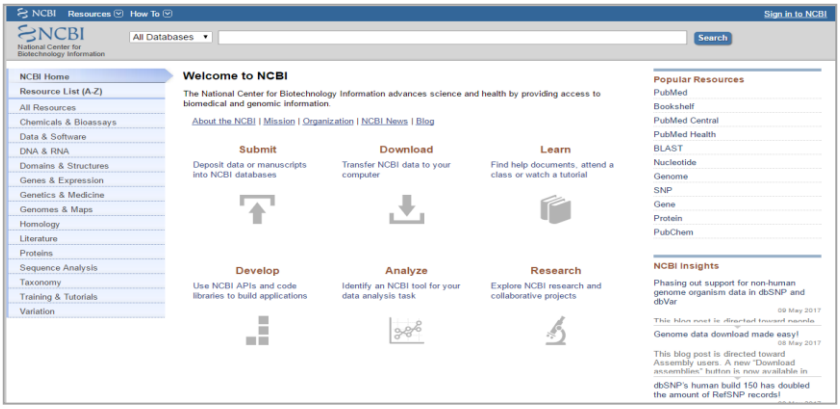


Figura 53. Pantalla de bienvenida del portal de NCBI

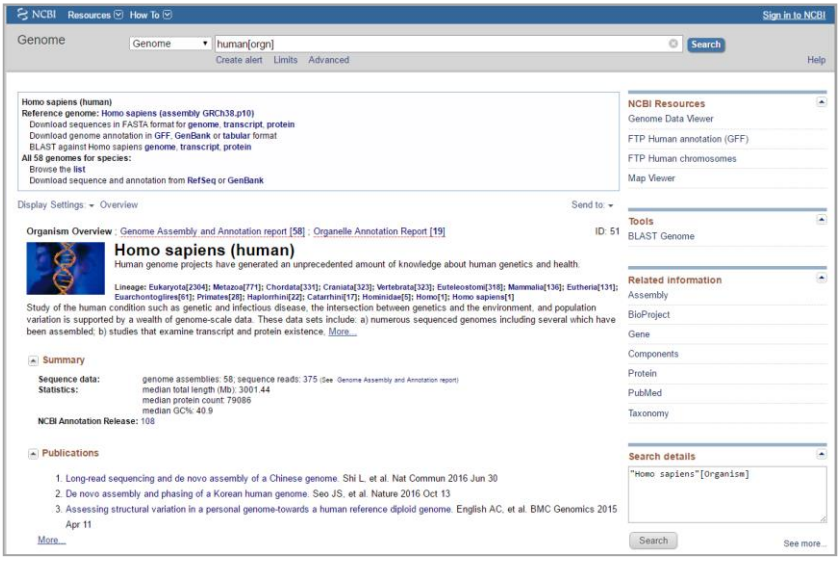


Figura 54. Búsqueda de información sobre el "*genoma humano*"

Las Figuras 53 y 54, presentan la pantalla de bienvenida del portal de NCBI y la búsqueda de información sobre "*Genomas*", y en este caso

específico sobre el genoma humano. En esta última explica toda la información sobre la versión del ensamblaje del genoma: *Homo sapiens* (human) → *assembly*: GRCh38.p10, así como unos datos estadísticos sobre las lecturas de la secuencia, GC%, las publicaciones asociadas, entre otros.

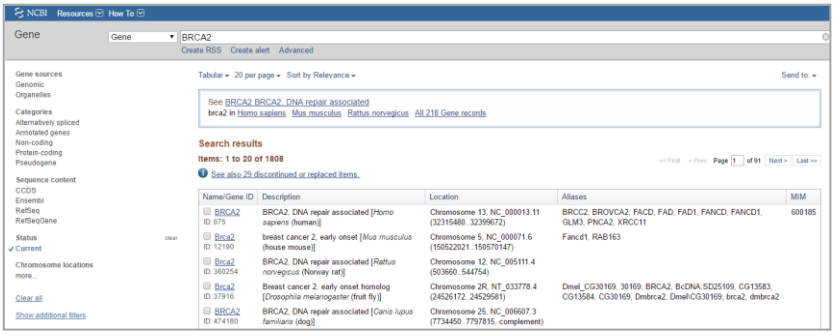


Figura 55. Búsqueda del *Gen* "*BRCA2*" en el portal de NCBI

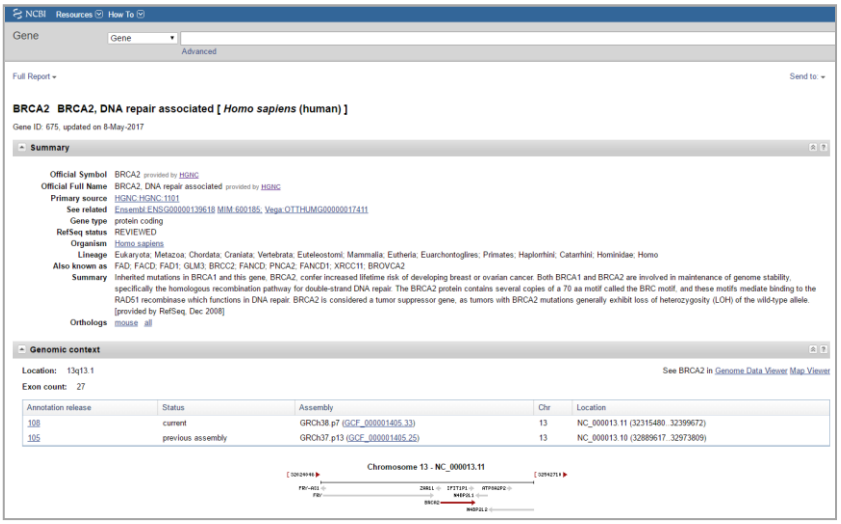


Figura 56. Información facilitada para el gen "*BRCA2*".

Las Figura 55 muestra el proceso de búsqueda de un gen especifico ("*BRCA2*") en el portal de NCBI, desde su apartado "*Gene*". La Figura 56 presenta la información facilitada por NCBI sobre el gen "*BRCA2*", como, por ejemplo, símbolo oficial, tipo de gen (*protein coding*), nombres sinónimos, entre otros.

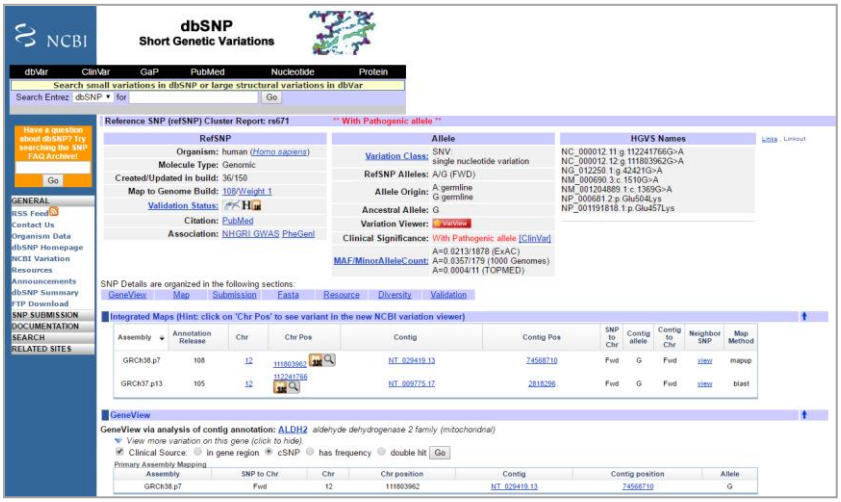


Figura 57. Búsqueda de la variación "*rs671*" en el portal de dbSNP

La Figura 57 presenta los resultados obtenidos en el portal de dbSNP para la variación "*rs671*", aquí se pueden observar las referencias del SNP (como, *organismo*, *tipo de molécula*, *validación*, *citación*, etc.), los detalles de los alelos (*alelo de referencia*, *alelo cambiado*, *significancia clínica*, etc.), información de la variación en la notación HGVS, entre otros.

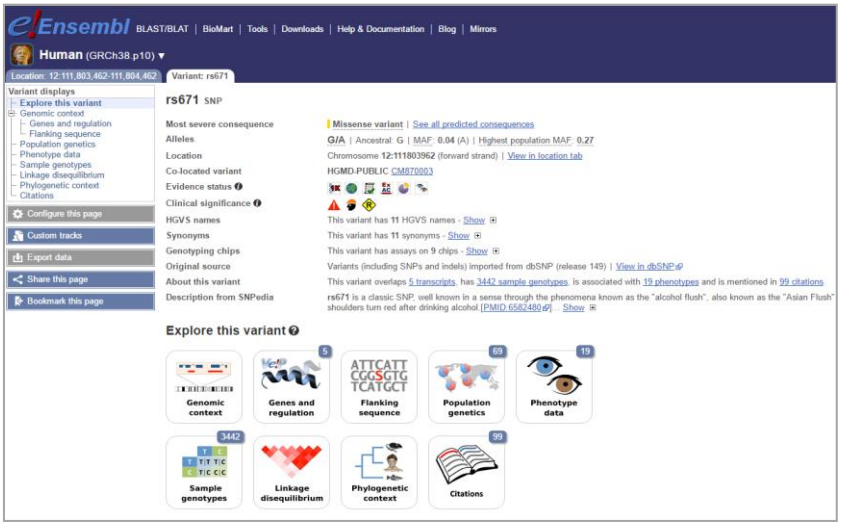


Figura 58. Resultado de búsqueda de la variación "*rs671*" en *Ensembl*

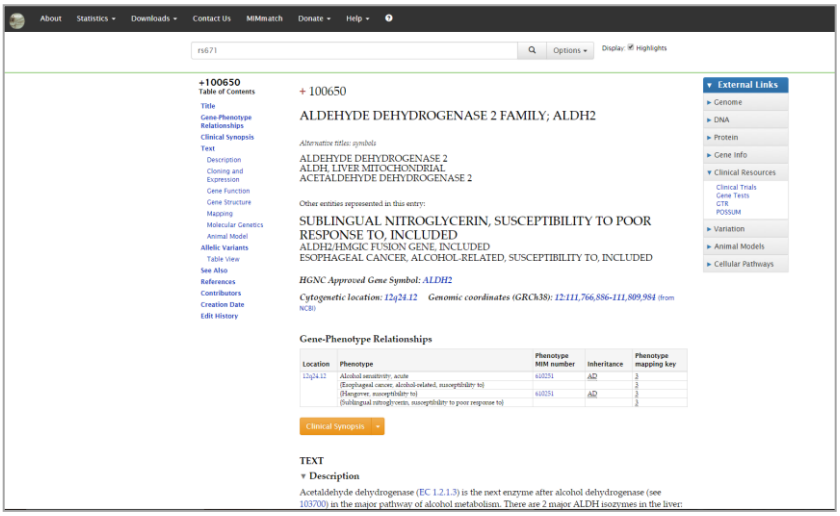


Figura 59. Búsqueda de la variación "*rs671*" en el portal de *OMIM*

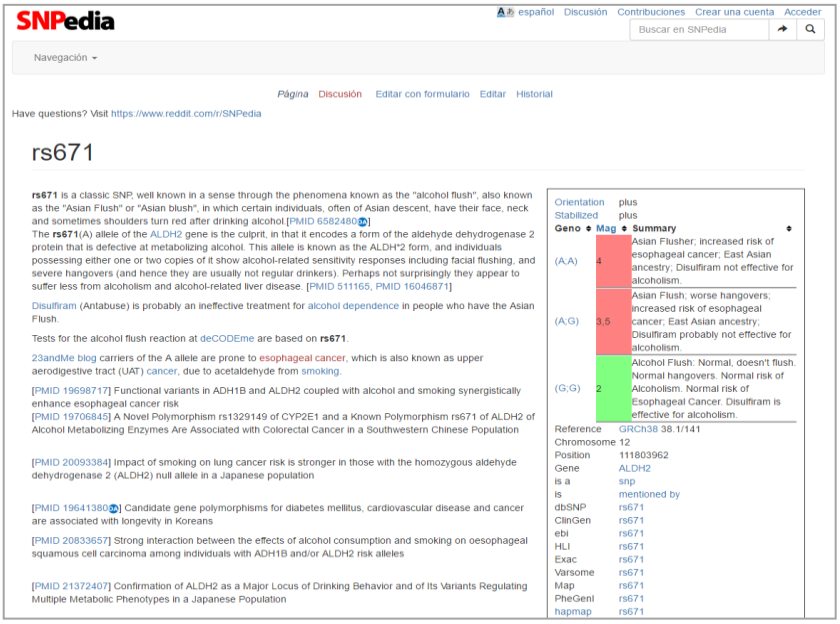


Figura 60. Búsqueda de la variación "*rs671*" en el portal de SNPedia

Las Figuras 58, 59 y 60 muestran la información relacionada con la variación "rs671", en los repositorios de *Ensembl*, *OMIM* y *SNPedia*.

Esta variación se encuentra en el gen "*ALDH2*", y la misma se asocia al fenotipo de la "*dependencia al alcohol*".

6.2 Base de Datos del Genoma Humano (HGDB)

En este trabajo se presenta el desarrollo de un *Sistema de Información Genómico* (GeIS), el cual tiene como objetivo las tareas de: *recolección*, *almacenamiento*, *gestión* y *distribución* de la información relacionada al comportamiento del genoma humano.

Este sistema de información genómico está basado en el *Modelo Conceptual del Genoma Humano* (MCGH) versión 2, descrito anteriormente en el Capítulo 4. La *Base de Datos del Genoma Humano* (HGDB) representa el conocimiento sobre variaciones y fenotipos cargados para la generación de diagnósticos genómicos. Para la transformación del MCGH definido al esquema de base de datos (*modelo lógico*) se utilizó la herramienta *Moskitt*[26] (Modeling Software KIT) permitiendo esta transformación de forma semiautomática.

El proyecto *Moskitt* tiene como objetivo proporcionar un conjunto de herramientas de código abierto, y una plataforma tecnológica para apoyar la ejecución de métodos de desarrollo de software basados en enfoques impulsados por modelos, la misma incluye: herramientas de modelado gráfico, transformaciones de modelos, generación de código, entre otros [169]. Es importante resaltar que en esta tarea se encontraron dos niveles distintos de abstracción del modelo, por lo que existen algunas diferencias entre ellas.

El modelo conceptual representa el dominio desde la perspectiva del conocimiento científico. Y, por otro lado, el esquema de base de datos se centra en el almacenamiento y recuperación de los datos de manera eficiente. Es por ello, que los detalles de la representación física deben ser considerados para mejorar la implementación final.

Es importante enfatizar la integración de dos tablas en el esquema de base de datos, "*Validation*" y "*Curator*". Estas tablas no forman parte de la representación del conocimiento del dominio, pero son necesarias para el desarrollo e implementación del sistema.

[26] *https://www.prodevelop.es/es/productos/moskitt*

Como se comentó en la sección anterior, los procesos de carga de la base de datos del genoma humano (HGDB) se han ejecutado de forma *masiva* y *selectiva*. Se han realizado pruebas de las dos maneras, pero el objetivo de esta tesis es desarrollar y explotar la carga selectiva de la información genómica. En la Figura 61 se muestra el esquema de base de datos implementado en este trabajo.

Figura 61. Esquema de Base de Datos (HGDB)

La descripción de las tablas que conforma la base de datos del genoma humano se encuentra definidas en el diccionario de datos desarrollado en el "*Anexo A*".

6.4.1 Selección de los repositorios de datos

Para la elección de los repositorios de datos, se plantearon varios requisitos, los cuales se planificaron con el fin de abordar distintas fases de carga. Las fases planteadas se clasificaron de la siguiente manera:

1. **Carga masiva de distintos repositorios, especial fenotipo de estudio "*cáncer de mama*".**
   Para esta primera fase se realizaron diversos estudios y análisis de los repositorios de datos genómicos disponibles, y se decidió implementar la carga del HGDB utilizando los siguientes repositorios: *NCBI*, *dbSNP*, *UMD* y *BIC*.

   NCBI (*National Center Biotechnology Information*) [170] es una de las principales fuentes de datos genómicas, y cuenta con datos curados sobre los conceptos estructurales de la secuenciación del ADN. A partir de este repositorio, se extrajo información sobre cromosomas, genes, transcripciones, exones y todo lo relacionado con la "*Vista Estructural*" del modelo conceptual del genoma humano (versión 2).

   En el caso de dbSNP (*Short genetic variations*) [71], BIC (*Breast Cancer Information Core*) [63] and UMD (*UMD Locus-Specific Databases*) [107], las tres almacenan información curada sobre las diferencias genéticas entre individuos.

   El principal motivo para utilizar *dbSNP* es la gran variedad de información que posee, pues otros repositorios sólo contienen las variaciones para un gen o una región especifica. Este repositorio de datos contiene las variaciones asociadas a todos los cromosomas, además, su información es actualizada constantemente.

   BIC y UMD fueron seleccionadas debido a los requerimientos de un proyecto de investigación paralelo, llamado "*Future Clinic*", que se estuvo desarrollando con otro equipo de investigación especializado en temas de "*cáncer de mama*" [171]. Este equipo colaboró con la validación del buen desempeño del *Sistema de Información Genómico* y la herramienta resultante (asociada).

2. **Carga selectiva, aplicación de la metodología SILE a enfermedades de origen genético.**
   En esta segunda fase de carga, se aplicó la metodología SILE a las siguientes enfermedades: (a) *Neuroblastoma*, (b) *Cataratas* y (c) *Alzheimer* como parte de la investigación y desarrollo de tres tesinas (grado, máster y doctoral, ver Figura 62).

   Como parte de este trabajo de investigación se guiaron estas tesinas como primera iteración de validación del HGDB sobre el cual estaría basado el prototipo propuesto.

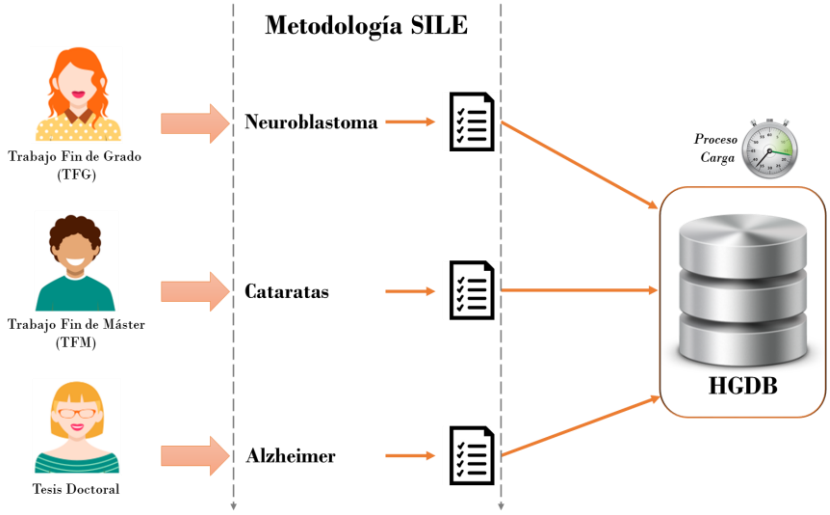


Figura 62. Propuesta carga selectiva

Dentro de los estudios realizados al implementar la metodología SILE, se evaluaron nuevos repositorios de datos. En los cuales se gestionaron nuevas estructuras de representación de los datos genómicos, lo que permitió enriquecer el conocimiento del dominio y sacar provecho de repositorios con estudios recientes y sólidos.

En el caso del Neuroblastoma se estudiaron los repositorios de *ClinVar* y *dbGap*. Para las cataratas se trató mediante *dbSNP* y bases de datos centradas en la enfermedad (NEI[27], *The*

[27] *https://nei.nih.gov/*

National Eye Institute), por último, en el tema del alzhéimer se estudiaron *ClinVar* y otras fuentes complementarias.

Estas fuentes de datos fueron estudiadas y los datos extraídos fueron mapeados y alineados conforme al modelo conceptual propuesto para el HGDB.

Para la carga de la información obtenida a través de los distintos repositorios genómicos se desarrollaron unos métodos o mecanismos de carga, los cuales son presentados en la Sección 6.4.2.

6.4.2 Módulo de carga (genética)

Para el proceso de carga de la HGDB se implementó un módulo de carga para almacenar los datos procedentes de los repositorios previamente comentados.

Este módulo de carga se desarrolló utilizando una estrategia ETL [172], con tres niveles diferentes: *Extracción-Transformación-Carga* (Figura 63). Cada nivel es completamente independiente el uno del otro, facilitando y clarificando el diseño del sistema (mejorando su *flexibilidad* y *escalabilidad*).

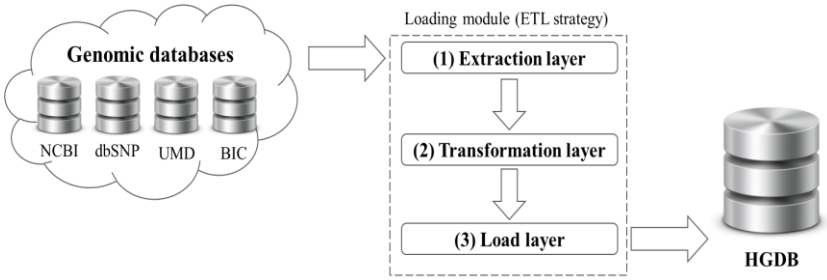


Figura 63. Módulo de carga

En la primera capa (1), se realizó la extracción de toda la información necesaria de los repositorios de datos origen. Dicha información no está estructurada, por lo que todos estos datos (en bruto) pasan a la segunda capa (2), en donde se realizan varias transformaciones con el objetivo de dar formato a los datos de acuerdo con la estructura del esquema de base de datos definido (HGDB). Estos datos

transformados se envían a la tercera capa (3), la cual se comunica directamente con la HGDB.

Dentro del grupo genoma del Centro PROS se han implementado mecanismos de carga, los cuales se definían dependiendo de la estructura y forma de obtención de los datos desde el repositorio origen (por ejemplo, *ficheros XML* o mediante un *servidor FTP*).

En el entorno bioinformático es frecuente encontrar desarrollos software basados en las tecnologías de Python (*https://www.python.org/*), esto debido a su mayor uso en la formación de expertos en el área bioinformática.

Como mecanismos iniciales de carga en el Centro PROS se desarrollaron "*scripts*" en Python con el objetivo de generar ficheros XML cuya estructura sea definida en base al MCGH, obteniendo de esta manera todos los genes almacenados en la base de datos de NCBI.

Para este trabajo se requería la generación de otros 5 ficheros (*gene_RefSeqGene.txt*, *output.xml*, *refseqgene1_genomic.gbff*, *refseqgene2_genomic.gbff* y *refseqgene3_genomic.gbff*, teniendo estos tres últimos el mismo formato). Con excepción de todos los ficheros, la generación del fichero "*output.xml*" consistía en la descarga de los datos mediante el siguiente sitio FTP: *ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/*. De igual forma se implementaron *scripts* para la extracción de información de la base de datos de BIC. En la Figura 64 se presenta un fragmento del código desarrollado para el *parser*[28] de BIC.

En este contexto se han estudiado y propuesto varios trabajos, los cuales buscaban facilitar el proceso de carga, teniendo en cuenta la complejidad (y gran desafío) para integrar las diversas estructuras de datos brindadas por los repositorios genómicos.

Otro enfoque planteado fue el desarrollo de un módulo de carga denominado "*Genoma Data Loader*", implementado con tecnologías *Java*. Esta aplicación integraba las tres capas del ETL (*extracción-transformación-carga*) y se centraba básicamente en la bases de datos de HGMD y BIC [173].

[28] *http://flanagan.ugr.es/xml/parser.htm*

```python
#!/usr/bin/env python

import sys, argparse, logging
from datetime import datetime

###################### MAIN ########################

if __name__ == '__main__':
    logging.basicConfig(filename=sys.argv[0]).rstrip('py').rstrip('.')+'.log',format='[%(asctime)s] %(levelname)s: %(message)s',\
                        filemode='w',level=logging.DEBUG)

    # define options
    parser = argparse.ArgumentParser()
    parser.add_argument('-i','--input', nargs='+', required=True, help='BIC data files, first BRCA1, second BRCA2')
    parser.add_argument('-o','--output', nargs=1, required=True, help='Output tabular data file')

    # parse args
    args = parser.parse_args()

    # retrieve options
    input = args.input
    output = args.output[0]

    var_index = {} # Variant Index

    # Interesting columns in BIC data files:
    #   0 - Accession number
    #   7 - HGVS cDNA
    #   13 - Clinically Important
    #   26 - Reference

    ng_identifiers = ["NG_005905.2", "NG_012772.3"]
    ng_point = 0
    for infile in input:
        with open(infile) as bic_file:
            # Leer primera linea (Cabecera)
            bic_file.readline()
            for line in bic_file:
                values = line.split('\t')
                # Eliminar posibles espacios en HGVS cDNA
                values[7] = values[7].replace(' ', '')

                # Check if the variation already exists in the variant index
                if var_index.has_key(values[7]) and (values[26].lower() not in ('-','unpublished')):
                    var_values = var_index.get(values[7])
                    var_values[3].append(values[26])  # Add a new reference for the variation
                    var_index[values[7]] = var_values # Update variant data
                else:
```

Figura 64. Trozo código Python: *parser BIC*

En el marco de esta Tesis Doctoral se desarrolló un prototipo de ETL para la carga automática de los datos (Figura 65), esto como parte (*co-dirección*) de la *Tesis de Máster de Manuel Navarrete Hidalgo* [174].

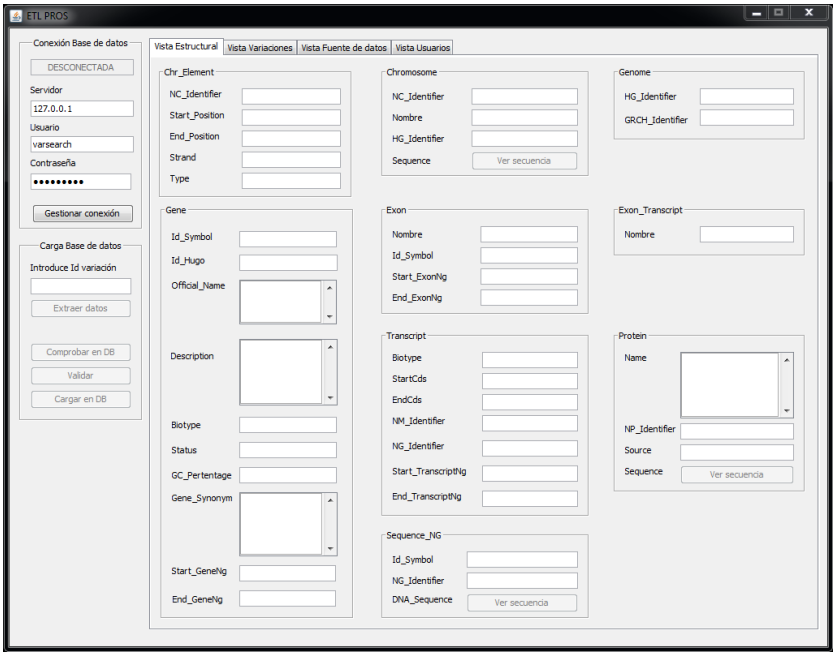


Figura 65. Ventana principal del prototipo *Software ETL*

La aplicación se inicia con el establecimiento de la conexión a la base de datos. Para ello, el prototipo dispone de una pequeña gestión de conexiones con la que se comprueba la comunicación directa con el servidor. Una vez establecida y validada, la aplicación informa del estado de la conexión y permite el proceso de *extracción* y *carga* de información desde los repositorios genómicos: *Clinvar*, *dbSNP*, *Gene*, *Nucleotide* y *Pudmeb*. La gestión de usuarios se realiza desde la "*Vista Usuarios*".

A continuación, se presentan algunas figuras del prototipo:

a) *detalle vista de variaciones* (Figura 66);
b) *detalle de extracción de la información* (Figura 67); y
c) *detalle de la vista de fuente de datos* (Figura 68).

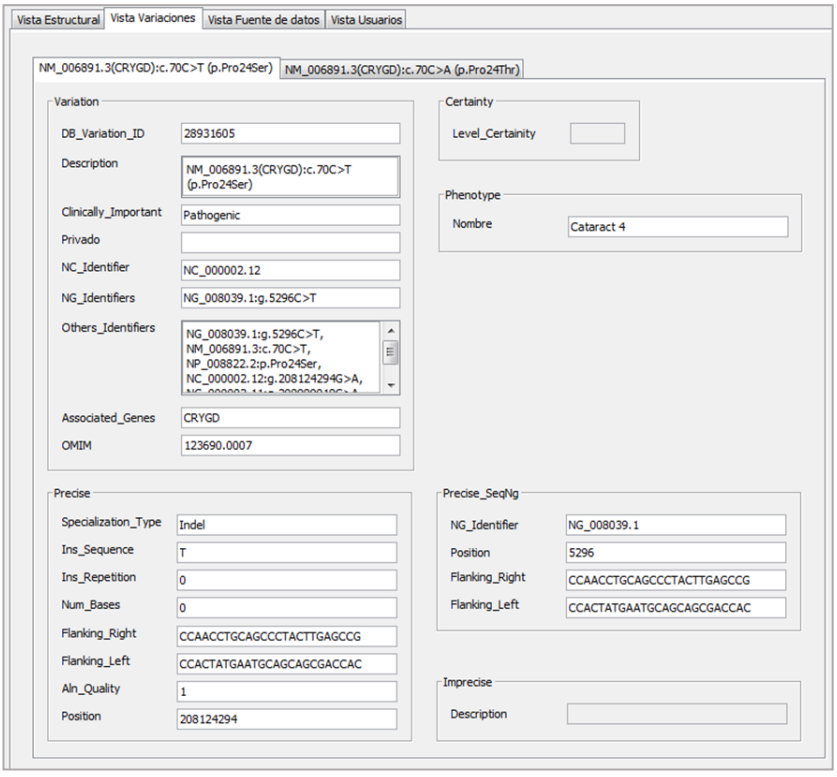


Figura 66. Detalle vista de variaciones

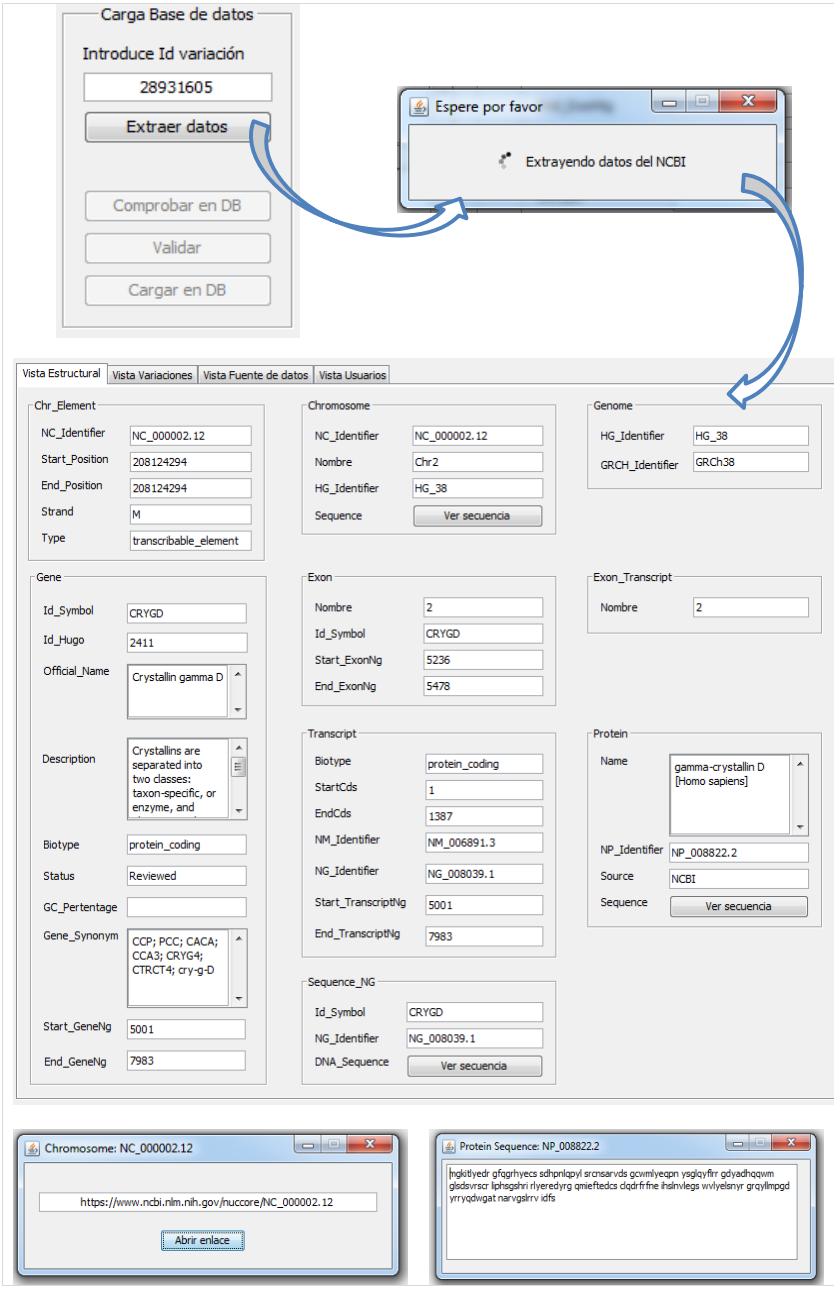


Figura 67. Detalle extracción de información (*Vista Estructural*)

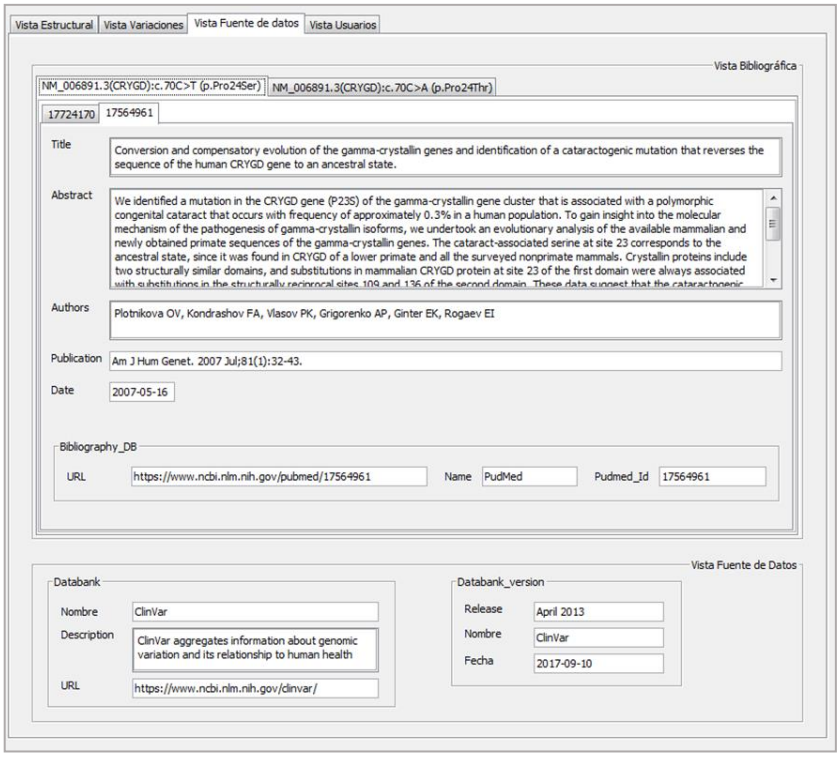


Figura 68. Detalle Vista Fuente de datos

Para más información sobre este módulo de carga y los datos cargados en la HGDB se puede consultar el trabajo titulado: "*Diseño e Implementación de un Sistema de Información Genómico para el Diagnóstico de la Catarata Congénita utilizando la Metodología SILE*" presentado por *Manuel Navarrete Hidalgo* (Septiembre 2017) en la Universitat Politècnica de València [174].

6.3 Ficheros VCF

Para la realización de las pruebas del prototipo se utilizarán ficheros VCF, los cuales representarán las "*muestras de pacientes*" para el análisis genómico (contrastando las variaciones del fichero con las cargadas en la HGDB).

El formato VCF ("*Variant Call Format*") es un formato genérico para almacenar datos de polimorfismo de ADN, tales como: SNPs,

inserciones, deleciones y variantes estructurales, junto con anotaciones opcionales derivadas de diferentes bases de datos.

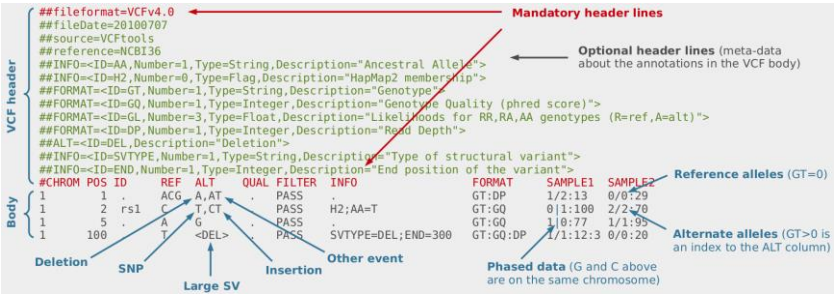


Figura 69. Estructura fichero VCF [175]

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G   A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T   A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A   G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T   .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Figura 70. Ejemplo fichero VCF [176]

Los VCF normalmente son almacenados de una forma comprimida y puede ser indexado para la recuperación rápida de datos de variantes de un rango de posiciones en la secuencia de referencia del genoma (en las Figuras 69 y 70 se muestra la estructura y un ejemplo del fichero VCF).

Este formato fue desarrollado para el *Proyecto 1000 Genomas*, y también ha sido adoptado por otros proyectos como: *UK10K*[29], *dbSNP* y el proyecto *NHLBI Exome*[30] [177].

[29] *https://www.uk10k.org/*

[30] *http://evs.gs.washington.edu/EVS/*

6.4 VarSearch (VS-prototipo)

El objetivo de esta sección es presentar el prototipo ("*VarSearch*") desarrollado como método de implementación y explotación del modelo conceptual propuesto. Hoy en día, los expertos en genómica realizan sus actividades con herramientas bioinformáticas (*software*) para la generación de diagnósticos genómicos, aunque la realidad práctica es que dichas soluciones no satisfacen plenamente sus necesidades.

Desde la perspectiva de los *Sistemas de Información* (SI), los principales problemas (reales) radican en la falta de aplicación de enfoques, es decir, técnicas de *Ingeniería de Software* que permitan generar estructuras correctas para la gestión de datos. Es por ello por lo que la comprensión del dominio genómico representa un gran desafío como ya se explicado anteriormente (*dispersión*, *heterogeneidad*, *inconsistencia* de los datos, etcétera).

Como solución, y para demostrar las ventajas del modelado conceptual en dominios complejos –*como la genómica*- se presenta "*VarSearch*", una herramienta basada en la web para la generación de diagnósticos genómicos, la cual incorpora el Modelo Conceptual del Genoma Humano (MCGH) y saca provecho de las últimas tecnologías de secuenciación, como, por ejemplo, NGS (*Next-Generation Sequencing*).

VarSearch es una aplicación web que permite el análisis de variaciones obtenidas de la secuenciación de ADN sobre muestras biológicas, y que son almacenados en archivos de formato FASTA [178] o VCF [177]. Esta aplicación permite el acceso con distintos roles (funciones) de usuario, con el objetivo de facilitar un espacio privado en el HGDB para que cada usuario pueda manejar sus propias variaciones. El espacio privado ofrece a los usuarios la inclusión de validaciones para variaciones que consideran o sugieren como "*relevantes*".

Además, permite el almacenamiento de las variaciones de los usuarios con el fin de encontrar similitudes en el proceso de análisis de los ficheros. Una ventaja proporcionada por los usuarios es la integración de la información obtenida de las distintas fuentes de datos junto con las validaciones realizadas por el usuario en la HGDB. Esta acción representa una mejora en el desempeño relacionado con la "*búsqueda de variaciones*".

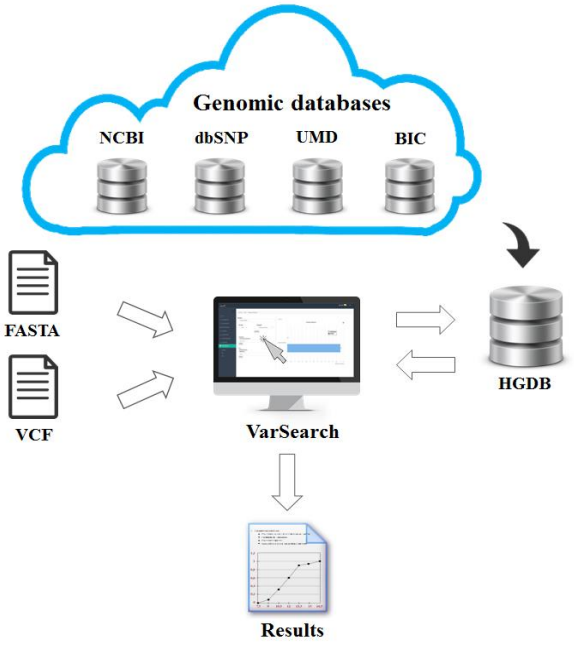


Figura 71. Aplicación VarSearch

VarSearch es capaz de encontrar variaciones en la HGDB a partir de un fichero facilitado (ver Figura 71). Las variaciones encontradas se muestran al usuario, calculando la información adicional que el fichero no posee y permite al usuario realizar validaciones sobre éstas. El usuario puede almacenar las variaciones del fichero que no se han encontrado en la HGDB. Después de insertar una o más variaciones no encontradas en un fichero (*porque son consideradas relevantes para el usuario*), al momento de volver a analizar dicho fichero, estas variaciones insertadas se encontrarán en la HGDB, por lo que el usuario podrá visualizarlas.

Con respecto a la funcionalidad de la herramienta, a continuación, se explica la agrupación representada en tres paquetes principales (Figura 72):

a) *Gestión de usuarios*: un usuario puede actuar como administrador y gestionar a otros usuarios, o puede crear nuevos usuarios y modificar o eliminar su información.
b) *Gestión de carga de datos*: el sistema permite al usuario cargar los ficheros a analizar, ya sea en formato VCF o FASTA, y

compara las variaciones de los ficheros contra las variaciones cargadas en la HGDB (que utiliza *VarSearch*).

c) *Análisis de datos*: tras finalizar el análisis y verificación de las variaciones en los ficheros de entrada, el usuario puede listar las variaciones y clasificarlas por múltiples criterios, como, por ejemplo, posición, cromosoma, existencia en la base de datos, entre otros).

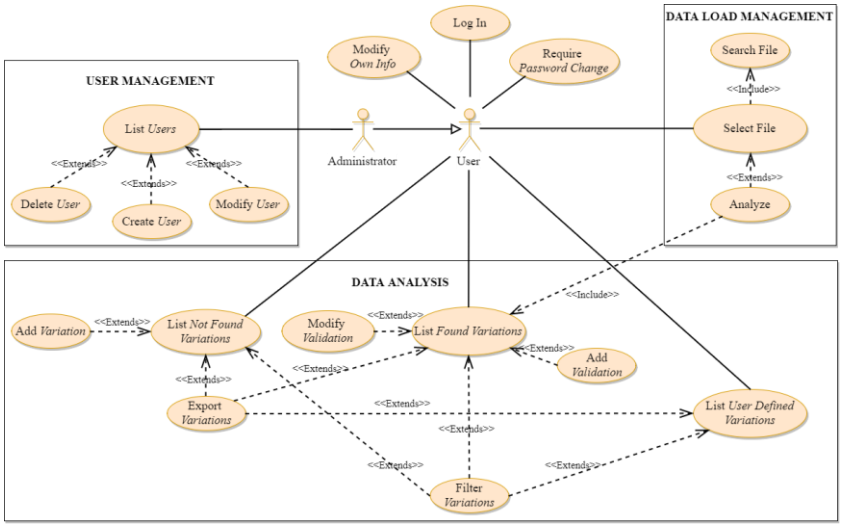


Figura 72. Diagrama de Caso de Uso General: *VarSearch*

También existen una serie de funcionalidades relacionadas con el control de sesiones y la modificación de la información sobre cuentas de usuario, las cuales no se han agrupado en ningún paquete de funcionalidad.

En la actualidad, la gestión optimizada de la información es uno de los recursos primarios de toda empresa u organización. Por esta razón, la privacidad de la información es un tema fundamental para tener en cuenta.

VarSearch permite la privacidad de la información. Cuando el usuario crea una validación en una variación, puede establecer la privacidad de la misma: (a) *contenido público*, si está dispuesto a compartir sus observaciones/anotaciones con otros usuarios, o (b) *contenido privado*, el cual sólo permite el acceso al usuario propietario.

6.4.1 Arquitectura de VarSearch

VarSearch está basado en un EGF (*E-Genomic Framework*), el cual está descrito en profundidad en los trabajos [163], [179] (ver Figura 73). Considerando en cuenta el ámbito de aplicación, y con el fin de hacerlo accesible a todos los usuarios, *VarSearch* ha sido desarrollado como una aplicación web.

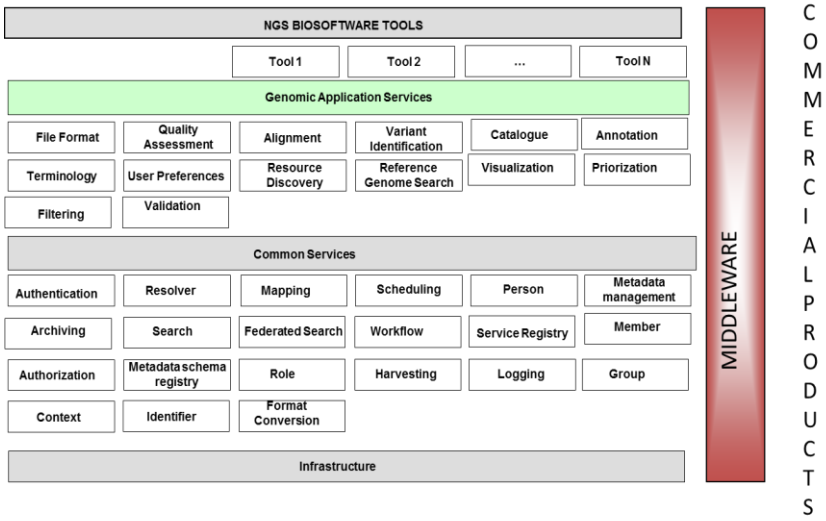


Figura 73. E-Genomic Framework y *VarSearch*

Dada la heterogeneidad de los navegadores actuales, *VarSearch* se ha implementado en un lenguaje común a todos ellos. Para asegurar la interoperabilidad, esta herramienta se ha desarrollado con tecnologías HTML5 [180]. Para la gestión de la información se generó una base de datos (HGDB) sobre MySQL (basada en el modelo conceptual).

La arquitectura de *VarSearch* consta de los siguientes elementos (ver Figura 74):

1. Una base de datos distribuible basada en MySQL.
   Para su implementación se utilizaron las siguientes herramientas software: *Navicat Enterprise* [181] y *MySQL Workbench*[31]. Para la validación (inicial) de la base de datos, se realizó la carga de la información relacionada con los cromosomas 13 y 22.

[31] *https://www.mysql.com/products/workbench/*

2. Un conjunto de servicios REST (*Representational State Transfer*) [182] [183] desarrollado en *Java* utilizando *Hibernate*[32] y *Jersey*[33], los cuales se despliegan en un servidor Tomcat 7[34].

3. Una aplicación web que utiliza el Framework Bootstrap[35] para la organización general de la interfaz y los archivos, junto con jQuery [184] para definir los componentes avanzados de la interfaz e invocar los servicios REST.

4. Se incluye también un "*mini*" servicio REST para gestionar los usuarios y roles, la cual se basa en la misma arquitectura y tecnologías que el resto de los servicios REST. La capa de datos se basa únicamente en MySQL.

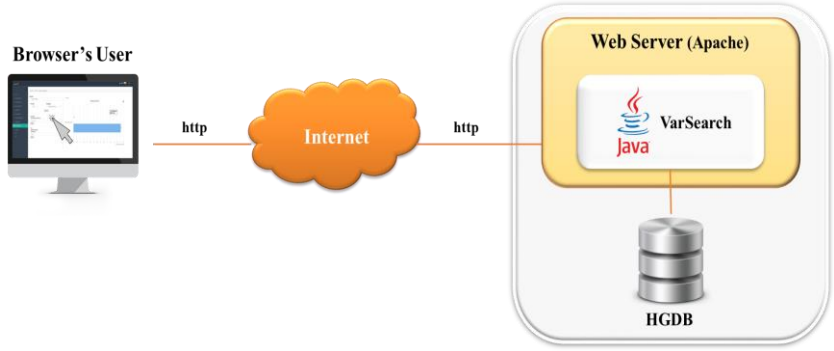


Figura 74. Arquitectura de *VarSearch*

[32] *http://hibernate.org/*
[33] *https://jersey.java.net/*
[34] *https://tomcat.apache.org/index.html*
[35] *http://getbootstrap.com/*

6.4.2 Guía de uso VS

El punto de entrada de la aplicación es un fichero con variaciones (detectadas) generadas por una máquina de secuenciación en formato VCF o FASTA. A partir de esta entrada, se busca en la HGDB para detectar las variaciones incluidas en el repositorio, facilitando información adicional sobre la enfermedad que produce y su bibliografía asociada. En esencia, los usuarios siguen estos pasos para trabajar con la herramienta:

1. Los usuarios acceden a la aplicación y proceden a cargar un fichero VCF/FASTA mediante el siguiente formulario (Figura 75):

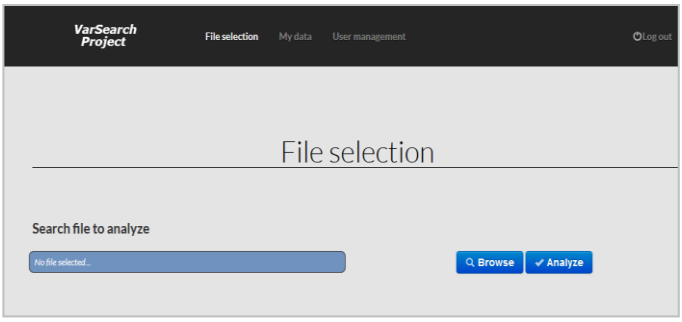


Figura 75. Selección y carga de fichero a analizar

2. Una vez cargado el fichero, el usuario pulsa el botón "*Analizar*", y el fichero se procesa como se muestra en la Figura 76.

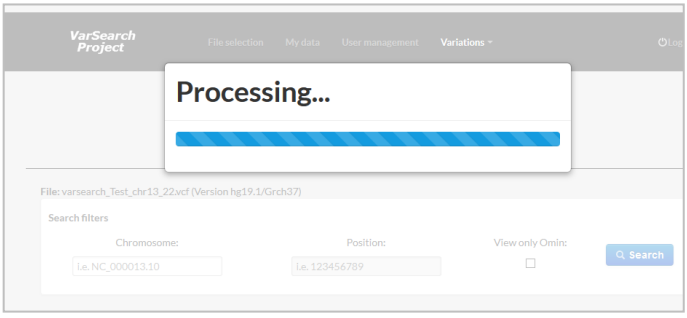


Figura 76. Tarea de análisis del fichero subido

Tras finalizar, se presentan todas las entradas en formato de tabla para cada entrada del fichero VCF/FASTA, se pueden consultar sus variantes (Figura 77).

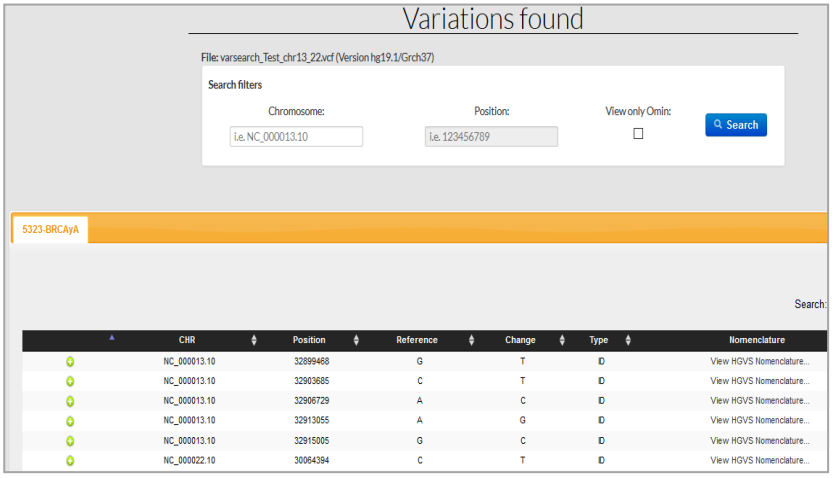


Figura 77. Variaciones encontradas en la HGDB

Para analizar el fichero VCF (en este caso), y anotar las variaciones (variantes), *VarSearch* se basa en las siguientes herramientas:

- *snpEff* [185]: es una herramienta de predicción de anotaciones y efectos. Anota y predice los efectos de las variantes en los genes (como, por ejemplo, los cambios de aminoácidos).
- *snpSift* [186]: consiste en una caja de herramientas que permite filtrar y manipular archivos anotados. *snpSift* ayuda a realizar la manipulación y filtrado de ficheros VCF en las etapas de –*pipelines*- de procesamiento de datos.

De esta forma, en lugar de reinventar la rueda, se han utilizado bibliotecas seguras y probadas. Por lo que se asegura un soporte estándar al VCF, mediante el uso de ficheros ANN[36] para la anotación de variaciones. Si se considerase útil otro tipo de información para la anotación, y no se encuentra cubierto en el procedimiento descrito, *VarSearch* utiliza el campo "*INFO*" para introducir los valores deseados.

[36] *https://fileinfo.com/extension/ann*

VarSearch se basa en un EGF[37] como se mencionó al principio de esta sección, por lo que los nuevos ficheros de anotación del genoma se pueden integrar rápidamente mediante el desarrollo del módulo analizador adecuado, tanto mediante un desarrollo personalizado o integrando una herramienta o biblioteca de terceros.

3. Finalmente, el usuario puede anotar las variaciones presentes en el fichero de entrada en la base de datos, y finalmente, descargar el fichero anotado (Figura 78) o visualizar su contenido en otra tabla *–listado de variaciones de usuario-* (Figura 79).

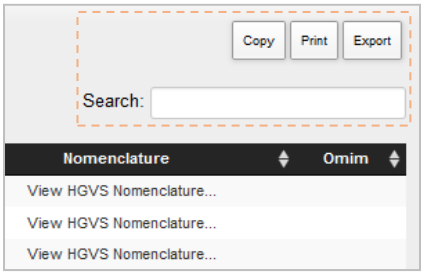


Figura 78. Barra de búsqueda/filtrado

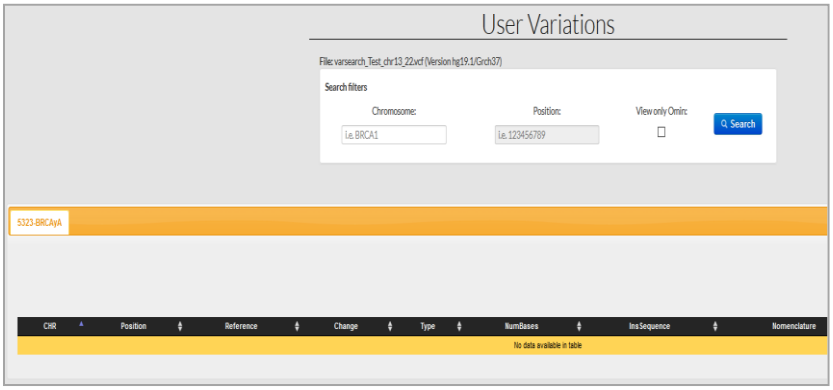


Figura 79. Listado de Variaciones de Usuario

[37] EGF: *E-Genomic Framework*

Los usuarios obtienen toda la información asociada con las variaciones encontradas en la HGDB, como, por ejemplo, *variaciones*, *fuentes de datos*, *significado clínico*, *cromosoma*, *posición dentro del cromosoma*, *alelo de referencia*, *alelo cambiado*, *tipo de cambio*, *nomenclatura HGVS* [114], *información de OMIM* [92].

Los datos obtenidos con la herramienta pueden ser filtrados a través de la barra de búsqueda. Además, se pueden *copiar*, *imprimir*, y *exportar* (en formatos: *.xls, *.csv y *.pdf) como se puede visualizar en la Figura 78. También es importante resaltar que, para las listas de variaciones, se pueden integrar "*validaciones de usuarios*" mediante la opción "*añadir validación*" (Figura 80), las cuales podrán ser consultadas en búsquedas futuras.

Figura 80. Formulario de inserción de validaciones en la HGDB

En la Figura 80 se muestra el formulario utilizado para la inserción de validaciones. De esta forma, se puede añadir algún *comentario* sobre la variación seleccionada por el usuario, así como asignar su significancia clínica, un identificador interno (dado por el usuario) y el grado de visibilidad de la validación (privada o pública).

Otra característica de *VarSearch* es la gestión de usuarios. Los nuevos usuarios se pueden crear y editar utilizando la opción "*Administración de Usuarios*" (Figura 81). Los atributos requeridos para gestionar el alta incluyen: *nombre de usuario*, *correo electrónico* y su *rol de acceso* (usuario general, administrador, etc.). En la pestaña "*Mis datos*" los

datos de usuario pueden ser actualizados (por ejemplo, la contraseña) por el administrador de la herramienta.

Uno de los objetivos de *VarSearch* es continuar su extensión e implementación sobre todo el conocimiento definido en el MCGH, ya que este prototipo se enfoca en explotar el conocimiento genómico sobre variaciones y fenotipos asociados (*vista de variaciones* del modelo conceptual). Algunos de los requerimientos a futuro se enfocan en el tratamiento de *Pathways* y *Rutas Metabólicas* [7].

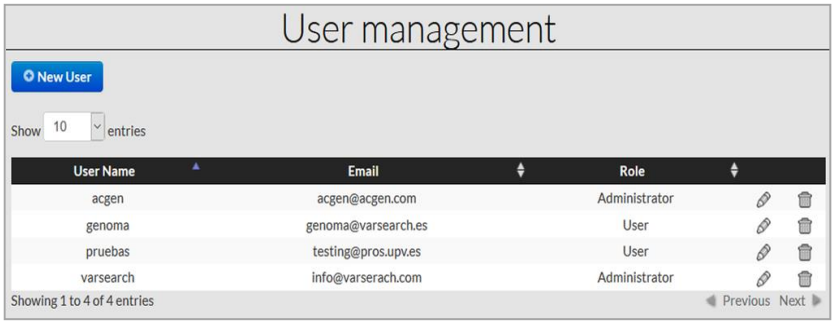


Figura 81. Gestión de usuarios

Esta herramienta facilita el análisis y búsqueda de variaciones, mejorando de esta manera la generación de diagnósticos genómicos asociados a enfermedades de origen genético. La aplicación web incorpora una amplia facilidad de uso para los usuarios finales, garantizándoles niveles de seguridad para sus datos [187].

6.4.3 Trabajos Relacionados

En este trabajo de investigación se consideraron las tres herramientas principales que buscan clasificar o anotar *variantes* (variaciones):

1) *SnpEff* [185]:
   Esta herramienta anota y predice los efectos de las variaciones genéticas, para esto construye una base de datos local que es cargada con información descargada de recursos/repositorios confiables. Tras finalizar el proceso de carga de la base de datos, *SnpEff* es capaz de analizar miles de variantes por segundo. Sin embargo, el proceso de carga de una base de datos es una tarea muy costosa desde el punto de vista de los

recursos, e incluso para su ejecución se recomienda aumentar los parámetros de memoria predeterminados por *Java*.

Por otra parte, a pesar de que esta herramienta se puede ejecutar de forma distribuida (utilizando Amazon Cloud Services) y ofrece interfaces web limitadas, la misma está orientada a la línea de comandos. SnpEff también puede ser integrado con otras herramientas como GATK[38] o Galaxy[39].

2) *Annovar* [188]:
Esta herramienta software se utiliza para anotar variantes. El primer paso para utilizar los scripts de *Annovar* consiste en llenar las tablas de la base de datos local usando un extenso conjunto de fuentes externas. El siguiente paso, consiste en anotar variantes de un fichero VCF para generar un fichero tabulado (personalizado) independiente.

wAnnovar (Web Annovar) basado en la web brinda un acceso fácil e intuitivo a las funcionalidades más populares de *Annovar*, esta herramienta permite a los usuarios enviar sus propios ficheros y esperar los resultados del informe de análisis realizado. Al igual que *SnpEff*, esta herramienta está orientada a la línea de comando, y no proporciona una API bien documentada para la integración de frameworks.

3) *VEP* (*Variant Effect Predictor*) [80]:
Esta herramienta se emplea para determinar el efecto de las variantes consultando bases de datos externas de forma directa, sin necesidad de cargar una base de datos en local (*aunque se recomienda por razones de rendimiento*).

Al igual que *SnpEff* y *Annovar*, *VEP* está orientada a la línea de comandos, y el acceso vía web es funcionalmente limitado. Para lograr la integración de esta herramienta con otras, se pueden ampliar sus funcionalidades básicas mediante una serie de complementos que ellos facilitan.

A continuación, se presenta la Tabla 9, la cual muestra una comparación entre *VarSearch* con estas tres herramientas:

[38] *https://software.broadinstitute.org/gatk/*
[39] *https://usegalaxy.org/*

Tabla 9. Comparación entre herramientas de anotación de variantes

Feature	**SnpEff**	**Annovar**	**VEP**	**VarSearch**
Distributed architecture	√	√	√	√
Type of application	Desktop	Desktop	Desktop	Web
Multiple database sources	√	√	√	√
Standard input formats	√	√	√	√
Standard output formats	√	X	√	√
Design paradigm	Data-oriented	Data-oriented	Data-oriented	Model-driven
Integration facilities	√	X	√	√

Como se muestra en la Tabla 9, *VarSearch* supera las limitaciones debido a que:

a) Está basado en una arquitectura *Java EE* para aplicaciones de varios niveles, el cual es un enfoque sólido para aplicaciones de alto nivel en entornos complejos y heterogéneos. Esto permite que *VarSearch* se pueda integrar fácilmente con otras aplicaciones web (el software está totalmente localizado), entre otros.
b) Se utiliza un framework orientado a servicios [179], el cual mejora aspectos de interoperabilidad e integración.
c) *VarSearch* utiliza una proyección del MCGH [112], por lo que depende de un paradigma orientado a *modelos* en lugar de un paradigma orientado a los *datos*.
d) Proporciona una interfaz web funcionalmente completa, con la capacidad de descargar resultados en formatos de fichero *–de salida estándar-*, que posteriormente pueden ser procesados por herramientas de terceros.

Como *VarSearch* sigue una arquitectura "*cliente/servidor*", la carga de datos no tiene ningún impacto en el rendimiento del cliente, mejorando de esta forma la experiencia del usuario. Además, la carga puede realizarse fuera de línea en el servidor para que los investigadores puedan consultar datos sobre la marcha con un tiempo de respuesta -*corto-*.

6.5 Caso de Estudio: Explotación del conocimiento genómico a través de VS

Para validar los resultados proporcionados por *VarSearch* se han realizado dos casos de estudio (*prueba*). El primero presenta el análisis de un fichero VCF utilizando la herramienta, explicando cómo se realiza el proceso y su funcionamiento. El segundo basado en la comparación del costo del tiempo dedicado para la búsqueda de variaciones: de forma *manual* y mediante *VarSearch*.

6.5.1 Explotación de tecnologías NGS

VarSearch es una aplicación restringida a ciertos tipos de usuarios, para acceder a la aplicación es necesario tener una cuenta de usuario facilitada por la empresa *Gembiosoft* [189]. Una vez que el usuario está conectado, el siguiente paso a realizar es la "*selección del fichero a analizar*".

Los formatos válidos para ejecutar el análisis son FASTA y VCF. De todos los formatos existentes hoy en día, estos dos son los más utilizados por los genetistas, por lo que su elección fue la más viable.

Una vez seleccionado el fichero deseado para analizar, en este caso un fichero VCF, *VarSearch* se encarga de leer todos los registros y transformarlos en variaciones.

Estas transformaciones mantienen la integridad del fichero: por ejemplo, los ficheros FASTA contienen una secuencia genética (*génica*), por lo que es la misma referencia para la variación "*NG*". Por el contrario, los ficheros VCF usan posiciones relativas a los cromosomas, que son representadas mediante el "*NC*". Una vez que los registros de los ficheros se han convertido en "*variaciones*", el siguiente paso es la búsqueda de estas variaciones en la base de datos (HGDB). Después de ejecutar el análisis, se pueden diferenciar entre las "*variaciones encontradas*" y las "*variaciones no encontradas*".

a) *Gestión de variaciones encontradas*:
   Las variaciones encontradas son las extraídas del fichero cuya información se ha encontrado en la base de datos del genoma humano. Esto significa que esta esta variación se ha encontrado en al menos un repositorio genómico. Una variación encontrada tiene mucha más información que la

variación obtenida del fichero, y permite calcular y enviar información más detallada al usuario.

Cuando finaliza el análisis del fichero VCF, el usuario obtiene todas las variaciones, presentando para cada una su notación HGVS, su identificador de origen de datos, significancia clínica, número de validaciones y las bases de datos con sus respectivas referencias bibliográficas. Dicha información se calcula para ambos formatos (VCF y FASTA); sin embargo, las variaciones VCF se ordenan por muestras, se puede visualizar la información en la Figura 82.

La Figura 82 presenta los resultados obtenidos analizando un fichero VCF con una sola muestra. Para la muestra '*5323-BRCAyA*' se han encontrado una serie de variaciones con su correspondiente información adicional, validaciones y referencias bibliográficos.

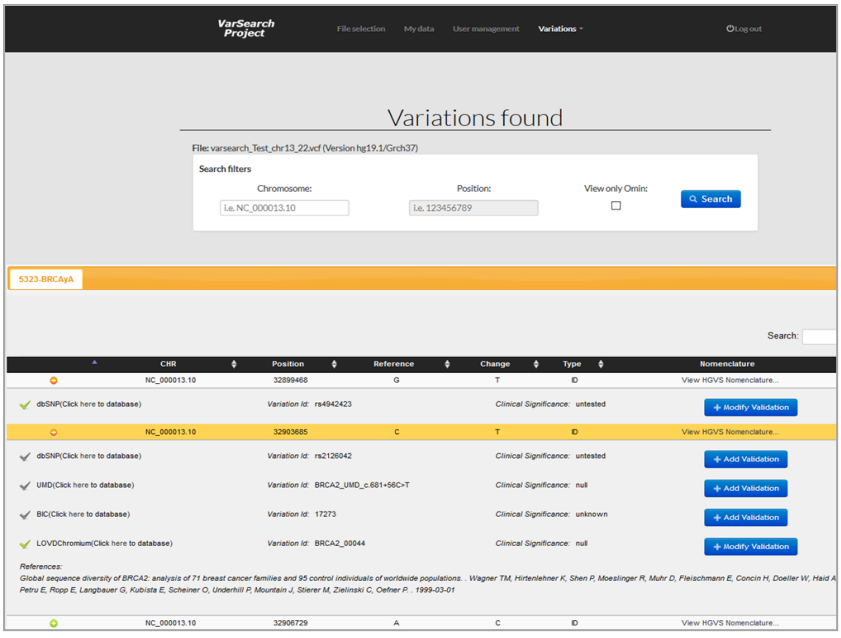


Figura 82. Lista de variaciones encontradas

Los usuarios pueden realizar validaciones sobre las variaciones. La columna de la "*validación*" corresponde al número de validaciones que posee cada variación. Un usuario sólo puede

realizar una validación en cada variación y esta puede ser pública o privada. Si la validación es privada, sólo el usuario que la ha creado podrá visualizarla.

Otra característica de *VarSearch* es el soporte de múltiples referencias bibliográficas. Una variación puede encontrarse en diferentes bases de datos y por ello contener distintas referencias bibliográficas. Teniendo en cuenta las múltiples referencias para una variación en una base de datos; para cada variación mostrada, se obtiene el repositorio de donde se ha obtenido la información, y para cada repositorio sus propias referencias bibliográficas.

b) *Inserciones y tratamiento de variaciones no encontradas*:
Una vez que se presentan las variaciones encontradas, puede ser el caso en el que el usuario que está procesando el fichero, este encuentre alguna variación relevante en el fichero y que no se haya encontrado en la base de datos (Figura 83). Basado en la experiencia y conocimiento del usuario, podría considerar algunas variaciones como relevantes a pesar de no ser encontrado.

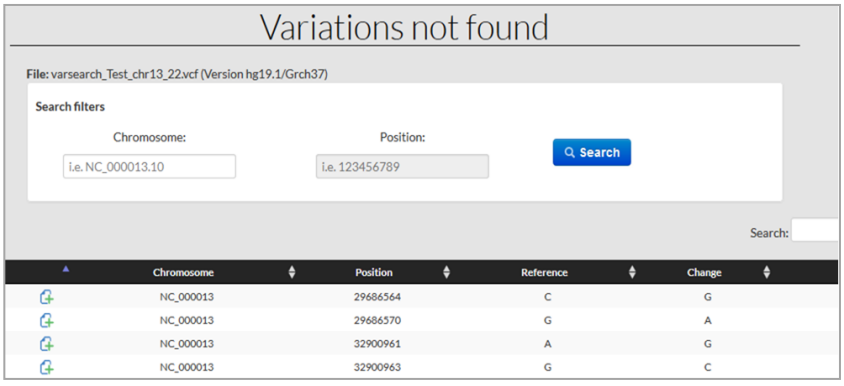


Figura 83. Lista de variaciones no encontradas

Por esta razón, *VarSearch* ofrece la posibilidad de *insertar* variaciones (como se muestra en la Figura 79). El usuario puede insertar las variaciones no encontradas o cualquier variación que considere clave para el estudio. Por lo tanto, si el usuario ha insertado un conjunto de variaciones que no se han encontrado, cuando se reanaliza el fichero, estas variaciones

insertadas se comparan con las variaciones en el fichero, mostrando de esta forma las similitudes entre ellas.

Con el objetivo de diferenciar las variaciones de los distintos repositorios y las creadas por los usuarios, el resultado que se obtiene se presenta por separado. Con esta separación, los resultados difieren según la experiencia del usuario y los resultados de años de estudio de las diferentes bases de datos genómicas.

6.5.2 Optimización del Tiempo

Para validar el efectividad y rendimiento del prototipo desarrollado, se han llevado a cabo varios experimentos para medir la optimización del tiempo en la búsqueda de variaciones.

Este estudio se ha realizado tomando en cuenta el tiempo dedicado a la búsqueda de variaciones de forma manual en comparación con la búsqueda automática de *VarSearch*, para todos los repositorios mencionados anteriormente.

La búsqueda manual de una variación implica la detección de la variación en el fichero VCF o FASTA, y la posterior búsqueda de la variación en las distintas bases de datos (*identificación* y *verificación* de la variación).

Para ver el rendimiento de *VarSearch* se ha realizado un experimento basado en la búsqueda de variaciones y su tiempo de coste asociado. El objetivo de este estudio es calcular la evolución del tiempo dependiendo del número de variaciones buscadas, y comparar los resultados obtenidos con *VarSearch* y la búsqueda manual. Para este propósito, el número de variaciones aumenta en '2', '5' y '7', y su tiempo de coste se calcula para cada uno de estos valores. Los resultados de este estudio se ven reflejados en la Figura 84.

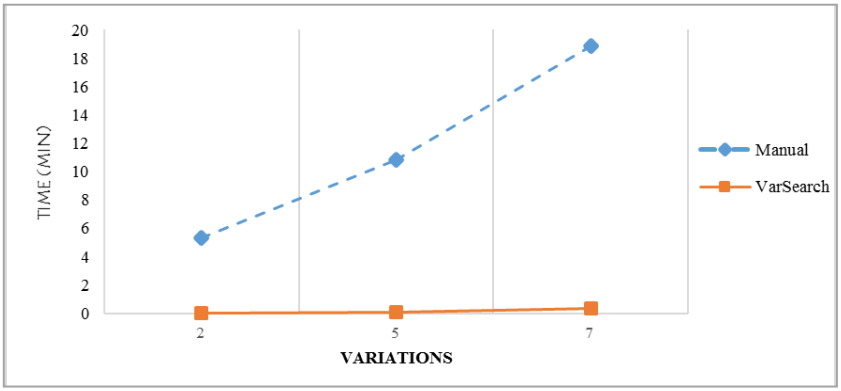


Figura 84. Optimización del tiempo

Como se puede visualizar en el gráfico (Figura 84), el costo de realizar una búsqueda manual se eleva a 5'32 minutos para 2 variaciones, 10'83 minutos para 5 variaciones y 18'89 minutos para 7 variaciones.

Sin embargo, para la búsqueda generada con *VarSearch* permanece constante entre 2 y 3 segundos para distintas variaciones. Mediante este estudio, se puede verificar el rendimiento facilitado por el prototipo.

La utilización de esta herramienta reduce significativamente el tiempo empleado en la búsqueda de variaciones. Una desventaja del proceso de búsqueda manual es que no se calcula información adicional para las variaciones. Y si esta información fuera necesaria, el tiempo de búsqueda aumentaría significativamente, sin embargo, con el uso de *VarSearch*, el tiempo se mantendría constante debido a que esta información ya está calculada en la búsqueda de variaciones.

6.6 GenesLove.Me: Canal de Interacción (VS y usuarios finales)

Esta sección presenta una visión global del *Proyecto GenesLove.Me* (GLM), el cual fue desarrollado como canal de interacción entre los usuarios finales (comunes) y los resultados obtenidos a través de *VarSearch*.

Como se ha explicado anteriormente, *VarSearch* tiene como objetivo la generación de diagnósticos genómicos. Esta herramienta está orientada específicamente para usuarios expertos, como, *genetistas*, *especialistas* o *laboratorios* que se dediquen a la búsqueda de variaciones para el diagnóstico de enfermedades. En el caso de *GenesLove.Me*, tiene como objetivo brindar *Test Genéticos Directos al Consumidor (TGDC)*, por lo que las enfermedades estudiadas anteriormente para los procesos de carga de la HGDB fueron ofertadas de forma directa a través de la aplicación web.

Este trabajo de investigación busca mejorar y contribuir a la *Medicina de Precisión* mediante el desarrollo e implementación de *Sistemas de Información Genómicos*. *GenesLove.Me* es una aplicación web desarrollada para proporcionar TGDC. Durante las etapas de análisis y diseño se plantearon los modelos de procesos de negocio (BPMN) y una representación conceptual (*modelado conceptual*), con el fin de mejorar los procesos involucrados en este tipo de servicio y proporcionar una plataforma basada en modelos para la gestión de los diagnósticos genéticos de una manera *escalable*, *segura* y *fiable*.

Los enfoques de *Ingeniería de Software* (IS) aplicados en el contexto genómico juegan un papel clave en el avance de la medicina de precisión o personalizada [37] [39].

La disponibilidad actual de los TGDC posee un gran número de ventajas para el dominio genómico, ya que facilita a los usuarios finales el acceso a los servicios de *diagnóstico temprano* de las enfermedades de origen genético.

Romeo-Malanda define los "*test genéticos directos al consumidor*", como un término que se utiliza para describir los servicios analíticos ofertados para la detección de polimorfismos y las variaciones genéticas relacionadas con la salud [190]. Aunque este tipo de análisis se encuentra disponible a través de sistemas de venta directa en

farmacias u organismos de atención de salud, *Internet* se ha convertido en el principal canal de distribución de los TGDC. El procedimiento habitual es tomar una muestra biológica en casa y enviarla a un laboratorio de análisis, y posteriormente los hallazgos (*resultados*) del análisis se comunican al cliente por teléfono, correo electrónico o mediante un portal de internet (acceso seguro) [191].

Al *GenesLove.Me* facilitar los diagnósticos genómicos generados a través de *VarSearch*, se garantiza que los datos utilizados han pasado procesos de estudio y validación (*metodología SILE*) antes de ejecutar la carga selectiva en la HGDB.

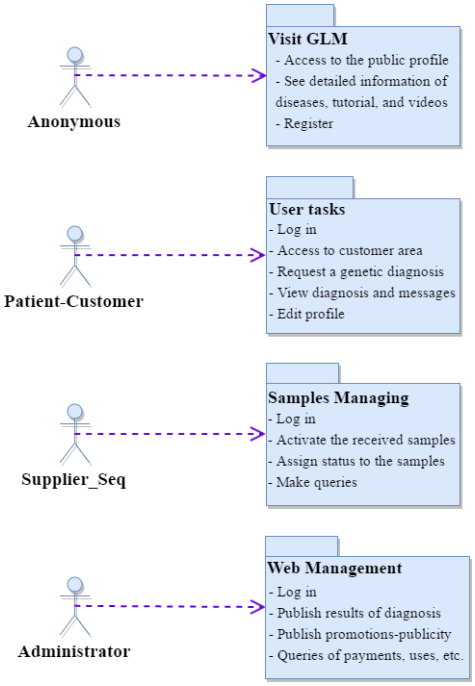


Figura 85. Diagrama de Paquete: *GenesLove.Me*

La Figura 85 presenta una vista general de la funcionalidad de *GenesLove.Me*, aquí se puede apreciar la participación de cuatro actores (usuario anónimo, paciente-cliente, suplidor de secuencia y el administrador) que son los que interactúan con la aplicación.

Los resultados del *Proyecto GenesLove.Me* se encuentran reportados en los siguientes trabajos: [127] y [164], en los cuales se puede consultar

más detalles sobre el BPMN y modelo conceptual planteado para la definición de este contexto "*Proceso del Diagnóstico Genómico*".

6.5.1 Arquitectura GenesLove.Me

GenesLove.Me es una aplicación web implementada bajo una arquitectura "*cliente/servidor*" como se muestra en la Figura 86. El lado del cliente, el navegador del usuario sirve como punto de interacción entre el usuario y la aplicación. El usuario final interactúa con una interfaz web gráfica fácil de usar, por la cual solicita los servicios (test genéticos) disponibles en la aplicación. Por otra parte, el lado del servidor se encuentra alojado en Internet, y contiene: a) el servidor web –*Apache 2.2*-, la parte lógica de la aplicación implementada con PHP, y b) la gestión de los datos mediante el motor de base de datos de MySQL 5.5.

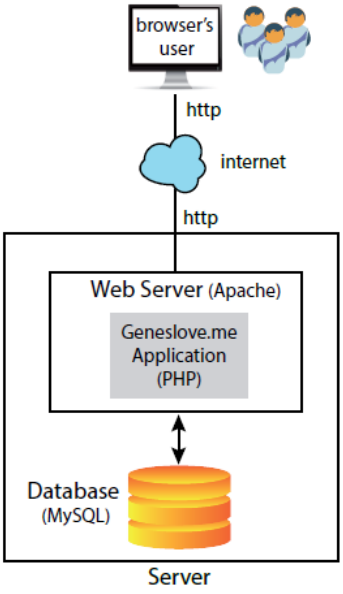


Figura 86. Arquitectura de *GenesLove.Me*

El diseño de *GenesLove.Me* permite a los clientes acceder a la gama de productos en línea (*pruebas clínicas*) desde cualquier lugar y en cualquier momento. De igual forma, los administradores de la gestión de la herramienta y tareas de negocio –*internas*- pueden acceder a través de una conexión a internet.

GenesLove.Me esta implementado sobre *Prestashop*[40], una plataforma CMS (Sistema de Gestión de Contenidos, de sus siglas en inglés "*Content Management System*") de código abierto para el comercio electrónico, el cual facilita la implementación de soluciones personalizadas orientadas a la comercialización de productos enmarcados en un proceso -*simple*- de compra-venta.

La plataforma incorpora módulos y plantillas de sitios webs para proporcionar, respectivamente, funcionalidad específica y estilo gráfico personalizado de acuerdo con las necesidades del negocio. El paquete predeterminado de *Prestashop* incluye módulos[41] de funcionalidad básica (como, por ejemplo, *clientes*, *productos*, *pedidos*, etcétera), los cuales son suficientes para crear y administrar una plataforma básica de comercio electrónico. Sin embargo, *Prestashop* permite incorporar módulos complejos de funcionalidad para adaptar los sitios web de acuerdo con las necesidades particulares, un ejemplo, es el módulo implementado para gestionar los procesos de pago. *GenesLove.Me* permite los siguientes métodos de pago: *transferencia bancaria*, *tarjeta de crédito*, *Paypal* o *cheques electrónicos*.

Con el objetivo de validar el proceso del diagnóstico genómico propuesto en este trabajo [164], se aplicó a los casos de prueba evaluados en la solución implementada. El escenario se basó en un grupo de cinco usuarios, quienes realizaron la solicitud del test genético "*Intolerancia a lactosa*" a través de la aplicación web (Figura 87).

Al iniciar el proceso, cada participante del estudio autorizó todo el procedimiento mediante la firma del "*consentimiento informado*" [192], el cual se convierte en el soporte legal para establecer los derechos y obligaciones del servicio entre ambas partes (cliente/empresa).

Los resultados correspondientes a los 5 participantes fueron entregados a través del portal –*GenesLove.Me*- tras finalizar las dos semanas de elaboración del diagnóstico genético (puesta en marcha de los procesos definidos en el modelo de BPMN).

[40] *https://www.prestashop.com*

[41] *http://addons.prestashop.com/en/2modules*

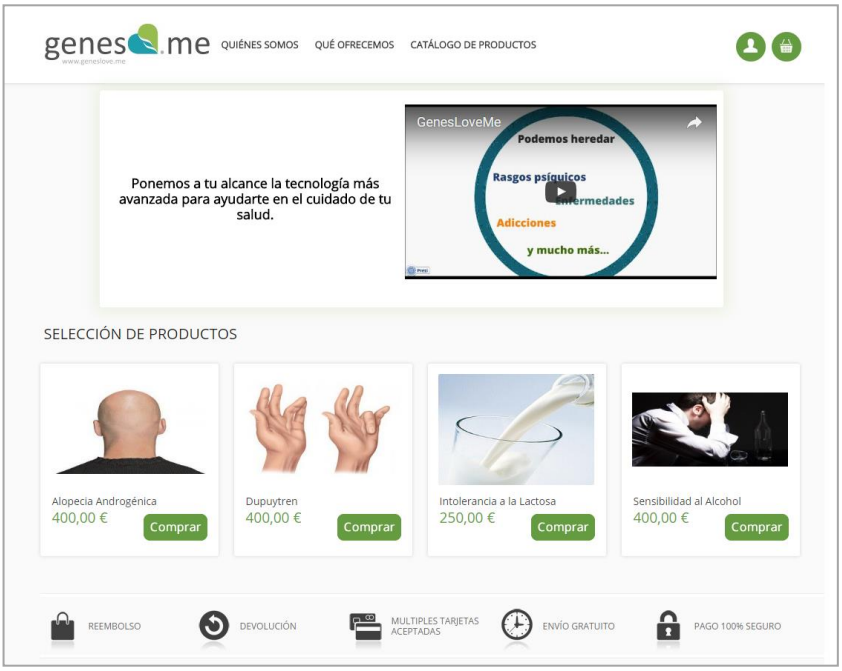


Figura 87. Página web de *GenesLove.Me* que muestra los TGDC disponibles (*Alopecia Androgénica*, *Intolerancia a la lactosa*, *Sensibilidad al alcohol*, *etc.*).

6.7 Conclusiones

VarSearch es un *framework* de análisis flexible que proporciona un poderoso recurso para explorar *variaciones genéticas*, tanto *codificantes* como *no codificantes*. Para ello, integra la entrada/salida del formato VCF con un conjunto en expansión de información genómica. Por lo tanto, este prototipo permite facilitar la investigación sobre las bases genéticas de las enfermedades humanas.

Hoy en día, las necesidades fundamentales de los laboratorios genéticos se han orientado en facilitar los procedimientos y tareas de los genetistas. El acceso a la web, la usabilidad y factibilidad, más la definición de los diferentes perfiles de esta herramienta, fueron los elementos claves para su desarrollo. Este conjunto de características permite al usuario configurar la herramienta *–de acuerdo con sus propias necesidades-*. Algunas de estas necesidades son: insertar variaciones genéticas y validar sus propias variaciones, para dar paso al aumento de su propio "*know-how*".

También es importante resaltar que integrando la oferta de *Test Genéticos Directos al Consumidor* (TGDC) a través de *GenesLove.Me*, se cubren todos los *stakeholders* involucrados en el proceso de generación de diagnósticos genómicos. Como se mencionó anteriormente con *VarSearch* se facilita una herramienta que gestiona los datos genómicos (información curada) de forma eficiente y eficaz, y que ese resultado sea transferido desde los genetistas o laboratorios clínicos a los usuarios finales, lo que permite potenciar y mejorar los métodos de tratamiento y prevención.

Este trabajo muestra como el diseño de *Sistemas de Información Genómicos* se ve mejorado y optimizado cuando sus bases están apoyadas en métodos dirigidos por modelos.

CAPÍTULO 7

Conclusiones

En este capítulo se presentan las conclusiones finales de la presente Tesis Doctoral. En primer orden, la Sección 7.1 explica de forma resumida las contribuciones principales obtenidas en el desarrollo de la Tesis Doctoral. A continuación, la sección 7.2 presenta el impacto de tesis, donde se muestran las distintas publicaciones académicas desarrolladas en el marco de la Tesis Doctoral (Sección 7.2.1), así como la colaboración en proyectos de investigación (Sección 7.2.2) y participación en la comunidad de modelado (Sección 7.2.3). Finalmente, en la Sección 7.3 se plantean las líneas de trabajo futuras.

7.1 Contribuciones principales

En el transcurso de la presente Tesis Doctoral, se ha justificado la necesidad de aplicar técnicas de modelado conceptual en el dominio genómico. El motivo fundamental se basa en la gran complejidad que define este contexto, el cual requiere del uso de enfoques y técnicas que permitan facilitar la gestión del conocimiento existente.

Gracias a los avances en las técnicas de secuenciación (NGS), como, por ejemplo, *mayor facilidad en secuenciación y reducción en el coste*[42], han dado lugar en los últimos años a la generación de grandes cantidades de información que debe ser gestionada de manera eficiente. Y, por otra parte, se encuentra la problemática de que no hay mucha gente utilizando las técnicas de modelado en este dominio.

La contribución esencial de esta Tesis Doctoral se basa en la caracterización de un *Modelo Conceptual holístico del Genoma Humano* como herramienta esencial para ese proceso de gestión efectiva y eficiente de datos genómicos. Mediante este modelo se logró mejorar el entendimiento del genoma humano con una representación gráfica (global) que permitió agrupar todos los elementos participantes en el comportamiento del genoma y sus interacciones (asociaciones).

Esta es la primera propuesta de un modelo conceptual holístico del genoma humano, que integra distintas fases o etapas del proceso biológico (por ejemplo, la definición de la estructura del genoma -*cromosomas y elementos del cromosoma*-, o de los procesos de -*transcripción*- para generar una proteína dada). En la consecución de este objetivo principal, se sumaron varias contribuciones, las cuales han sido plasmadas en este trabajo. Específicamente, las contribuciones de esta tesis han sido las siguientes (*cada punto asociado a un subobjetivo*):

1. La **definición y formalización de un Modelo Conceptual holístico del Genoma Humano**, el cual permitió alcanzar un mejor entendimiento del dominio. Además, de aportar una definición integral del genoma. Este modelo facilita una estructura basada en el conocimiento actual, y deja una puerta abierta para futuras extensiones y mejoras del MCGH. Mediante la utilización de este modelo conceptual del genoma humano se pretende armonizar la gran cantidad de datos existentes en el dominio genómico.

2. El **análisis y evaluación de la evolución del modelo conceptual**, el cual permite demostrar que en un entorno de continua evolución -*como es el genómico*- es importante estudiar la

[42] Actualmente, los costes de los *tests genéticos* facilitados por la empresa *23andMe* se encuentra en: $99 dólares (servicio de información ancestral) y $199 dólares (para el servicio de salud -incluyendo la información ancestral-), *https://www.23andme.com/*.

adaptación del modelo conceptual sobre los nuevos datos genómicos que se están manipulando en el día a día. Este análisis permitió establecer un conjunto de decisiones apoyadas por expertos, las cuales orientaban el modelo a una mejor forma de representación (*holística*) que satisfacía las necesidades tanto desde la perspectiva biológica como informática.

3. La **extensión del modelo propuesto** demostró la escalabilidad del MCGH. Esta propuesta afirma la necesidad continua que requiere un dominio como este, en donde es importante la manipulación (correcta) de la información existente para potenciar a mayor escala la medicina de precisión. La integración de *Haplotipos* en el MCGH es un ejemplo práctico de la capacidad de evolución conservando su definición inicial.

4. El **desarrollo de un prototipo basado en el modelo conceptual para la gestión de datos genómicos**, con el objetivo de facilitar un diagnóstico genómico *-precoz-* generado a partir de la base de datos del genoma humano (HGDB). Este paso incluyó la aplicación de una metodología (*sistemática*) para la obtención de los datos, por lo que se pasó de una "*carga masiva*" a una "*carga selectiva*" de datos. De esta manera se garantizaba la utilización de datos curados (*validados*) en el proceso de análisis del prototipo. A través de este prototipo los genetistas o laboratorios clínicos pueden gestionar sus datos, y llevar un control de las variaciones encontradas y no encontradas en las muestras analizadas. Finalmente, se plantea *GenesLove.Me* como canal de distribución para los test genéticos generados a partir de *VarSearch*, lo que permite concluir el proceso en manos del usuario final.

7.2 Impacto de la tesis

Este trabajo de investigación ha sido validado mediante la publicación de resultados en distintos foros académicos -*relevantes*- (a nivel nacional como internacional). Estos trabajos han sido evaluados desde las áreas de *Sistemas de Información* e *Ingeniería de Software*, como también en el contexto *Bioinformático y/o Biotecnológico* (ER, ENASE, RCIS, CLEi, Bioinformatics, entre otros). En la Tabla 10 se presenta un resumen de las publicaciones, y muestra la relevancia de cada una de ellas, indicando el tipo de comunicación. En esta tabla también se muestra el contenido de la publicación, indicando cuales capítulos de la tesis han sido cubiertos en la contribución.

La trayectoria durante este camino doctoral ha sido de gran satisfacción y crecimiento (*a nivel personal como profesional*), poniendo en manifiesto las habilidades adquiridas mediante la colaboración en proyectos de investigación, co-dirección de proyectos (*tesinas*), organización de conferencias científicas, la obtención de varios premios relevantes (asociados a la investigación), entre otros.

7.2.1 Publicaciones

Publicaciones en Revistas (1):

[1] **J. F. Reyes R.**, O. Pastor, F. Valverde and D. Roldán, "How to deal with Haplotypes data: An Extension to the Conceptual Schema of the Human Genome", ***CLEI electronic journal***, vol. 19, no. 3, paper 2, 2016. DOI: http://dx.doi.org/10.19153/cleiej.19.3.2.

Capítulos de Libro (3):

[2] **J. F. Reyes R.**, C. Iñiguez-Jarrín, and O. Pastor, "Genomic Tools*: Web-applications based on Conceptual Models for the Genomic Diagnosis", *selected papers from ENASE 2017 in Communications in Computer and Information Science (CCIS)*, Springer, 2017 (*to appear*).

[3] C. Iñiguez-Jarrín, A. García S., **J. F. Reyes R.** and O. Pastor, "Guidelines for Designing User Interfaces to Analyze Genetic Data. Case of Study: GenDomus", *selected papers from ENASE 2017 in Communications in Computer and Information Science (CCIS)*, Springer, 2017 (*to appear*).

[4] O. Pastor, A. León Palacio, **J. F. Reyes R.** and J. C. Casamayor, "Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome", *Conceptual Modeling Perspectives*, Verlag: Springer International Publishing, ISBN: 978-3-319-67270-0 / Electronic ISBN: 978-3-319-67271-7, pp. 25-40, October 2017. DOI: 10.1007/978-3-319-67271-7_3

Conferencias Internacionales (8):

[5] V. Burriel, **J. F. Reyes R.**, A. Heredia C., C. Iñiguez-Jarrín and A. León Palacio, "GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma", *IEEE 11th International Conference on Research Challenges in Information Science (RCIS 2017)*, pages 451-452, Brighton, UK, May 10-12, 2017. DOI: 10.1109/RCIS.2017.7956581 CORE B

[6] **J. F. Reyes R.**, C. Iñiguez-Jarrín and O. Pastor, "GenesLove.Me: A Model-based Web-application for Direct-to-consumer Genetic Tests", *The 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017)*, pages 133-143, Porto, Portugal, April 28-29, 2017. ISBN: 978-989-758-250-9. DOI: 10.5220/0006340201330143 CORE B

[7] C. Iñiguez-Jarrín, A. García S., **J. F. Reyes R.** and O. Pastor, "GenDomus: Interactive and Collaboration Mechanisms for Diagnosing Genetic Diseases", *The 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017)*, pages 91-102, Porto, Portugal, April 28-29, 2017. ISBN: 978-989-758-250-9. DOI: 10.5220/0006324000910102 CORE B

[8] **J. F. Reyes R.**, A. León Palacio and O. Pastor, "Software Engineering and Genomics: The Two Sides of the Same Coin?", *The 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017)*, pages 301-307, Porto, Portugal, April 28-29, 2017. ISBN: 978-989-758-250-9. DOI: 10.5220/0006368203010307 CORE B

[9] **J. F. Reyes R.**, O. Pastor, J. C. Casamayor and F. Valverde, "Applying Conceptual Modeling to Better Understand the Human Genome", *The 35th International Conference on Conceptual Modeling (ER2016)*, pages 404-412, Gifu, Japan, November 14-17, 2016. DOI: 10.1007/978-3-319-46397-1_31 CORE A

[10] **J. F. Reyes R.**, O. Pastor, F. Valverde and D. Roldán, "Including haplotypes treatment in a Genomic Information Systems

Management", *Proceedings of the XIX Ibero-American Conference on Software Engineering (CIbSE 2016)*, pages 1-14, Quito, Ecuador, April 27-29, 2016. ISBN: 9781510827189.

[11] **J. F. Reyes R.** and O. Pastor, "Use of GeIS for Early Diagnosis of Alcohol Sensitivity", *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016), Volume 3: BIOINFORMATICS*, pages 284-289, 2016. DOI: 10.5220/0005822902840289

[12] D. Roldán, O. Pastor and **J. F. Reyes R.**, "E-Genomic Framework for delivering genomic services. An application to JABAWS", *IEEE Ninth International Conference on Research Challenges in Information Science (RCIS 2015)*, pages 516-517, Athens, Greece, May 13-15, 2015. DOI: http://dx.doi.org/10.1109/RCIS.2015.7128915 CORE B

Workshops Internacional (1):

[13] A. León, **J. F. Reyes R.**, V. Burriel and F. Valverde, "Data Quality Problems When Integrating Genomic Information", *3rd. Workshop Quality of Models and Models of Quality (QMMQ 2016)*, in conjunction with the 35th International Conference on Conceptual Modeling (ER2016), pages 173-182, Gifu, Japan, November 14-17, 2016. DOI: 10.1007/978-3-319-47717-6_15 CORE A

Reporte Técnico (1):

[14] O. Pastor, **J. F. Reyes Román** and F. Valverde, "Conceptual schema of the human genome (CSHG)", *Technical Report*, Valencia, Spain, July 07, 2016. http://hdl.handle.net/10251/67297.

Aceptados para Publicación (2):

- ✓ M. Navarrete-Hidalgo, **J. F. Reyes Román** and O. Pastor López, "Design and Implementation of a GeIS for the Genomic Diagnosis using the SILE Methodology. Case Study: Congenital Cataract", *ENASE 2018*, Funchal - Madeira, Portugal, March 23-24, 2018.

- ✓ **J. F. Reyes Román**, D. Roldán M., A. García S., U. Rueda Molina and O. Pastor López, "VarSearch: annotating variations using an e-Genomics Framework", *ENASE 2018*, Funchal - Madeira, Portugal, March 23-24, 2018.

Tabla 10. Publicaciones realizadas en el marco de la Tesis Doctoral

Comunicación	Relevancia			Contribución					
	Tipo de Comunicación	Internacional	Ranking (CORE)	1. Motivación	2. Dominio Genómico	Estado del Arte	4. Evol. MCGH	5. Integración Haplotipos	6. Implementación
[4]	*Book Chapter (Springer)*	✓	-	✓	✓	✓	✓	-	-
[1]	*Journal*	✓	-	✓	✓	✓	-	✓	-
[8]	*Position Paper*	✓	[B]	✓	✓	✓	✓	-	-
[6] [7] [9] [10]	*Regular/Full Paper*	✓	[B][B][A][-]	✓	✓	✓	✓	✓	✓
[2], [3]	*Selected Paper for Book Chapter (Springer)*	✓	-	✓	-	✓	-	-	✓
[5] [11] [12]	*Short Paper*	✓	[B][-][B]	✓	✓	-	-	-	✓
[14]	*Technical Report*	-	-	-	-	-	✓	-	-
[13]	*Workshop Paper*	✓	[A]	✓	✓	-	✓	-	-

7.2.2 Proyectos académicos

Colaboración Proyecto DATAME
Un Método de producción de software dirigido por modelos para el desarrollo de aplicaciones Big Data. Ministerio de Economía y Competitividad del gobierno de España, Ref. TIN2016-80811-P. (Duración Proyecto: 30/12/2016 – 29/12/2020)

Colaboración Proyecto IDEO
Innovative services for Digital Enterprises with ORCA (Servicios Innovadores para Empresas Digitales con ORCA). Generalitat Valenciana, Ref. PROMETEO/2014/039. (Duración Proyecto: 01/01/2014 – 31/12/2017)

Colaboración Proyecto Accelerate
Incorporación de técnicas avanzadas de modelado para dar soporte a la aceleración de la innovación. Planet Media Studios, S.L., Ref. ITEA2 n.12014. 01/01/2015 - 20/07/2016. (79.061,00 €)

Colaboración Proyecto CAP
Collaborative Analytic Platform. Instituto de Medicina Genómica, S.L., Ref. ITEA2 n. 12010. 01/01/2014 - 18/05/2016. (30.000,00 €)

7.2.3 Participación en la comunidad de modelado

Tesis y asesor de proyectos:

- *Manuel Navarrete Hidalgo*. "Diseño e Implementación de un Sistema de Información Genómico para el Diagnóstico de la Catarata Congénita utilizando la Metodología SILE". Trabajo de Fin de Máster (TFM) - Máster en Ingeniería y Tecnología de Sistemas Software, Universitat Politècnica de València (UPV). Valencia, España. Septiembre, 2017. Co-dirección con Prof. Dr. Óscar Pastor L (9.0/10).
- *Clara Soler Pellicer*. "Diseño de un Sistema de Informacion Genómica para el Diagnóstico del Neuroblastoma". Trabajo de Fin de Grado (TFG) – Grado de Ingeniería Biomedica, Universitat Politècnica de València (UPV). Valencia, España. Julio, 2017. Co-dirección con Prof. Dr. Óscar Pastor L (9.3/10).
- *Miguel Ángel Moreno Molina*. "Proyecto Dupuytren: Informe Genético". Memoria de Prácticas Externas de Biotecnología, Centro PROS. Valencia, España. 2014. Co-dirección con Prof. Dr. Óscar Pastor L.
- *Itziar Ainhoa Sánchez López*. "Análisis y Desarrollo de Sistemas de Información y Herramientas Bioinformáticas para Técnicas de Secuenciación Genética de Nueva Generación". Proyecto Beca de Colaboración, Centro PROS. Valencia, España. 2014. Co-dirección con Prof. Dr. Óscar Pastor L.

Colaboración en docencia (prácticas):

- Asignatura del Máster Universitario en Ingeniería Biomédica (MUIB), Escuela Técnica Superior de Ingenieros Industriales (ETSII), UPV.
  - *Analysis of Genomic Data* (Curso: 2017-2018)
- Asignatura del Máster Universitario en Ingeniería y Tecnología de Sistemas Software (MITSS), Dpto. de Sistemas Informáticos y Computación (DSIC), UPV.
  - Sistemas de Información Aplicados a la Bioinformática: Gestión de Datos Genómicos (2015-2017)
- Asignatura del Grado en Ingeniería Biomédica, Escuela Técnica Superior de Ingenieros Industriales (ETSII), UPV.
  - El papel del Ingeniero Biomédico (Curso: 2015-2017)

Organización de conferencias científicas:

- Miembro del Comité Organizador - The 36th International Conference on Conceptual Modeling (**ER2017**). Valencia, España, noviembre 6-9, 2017.
- Miembro del Comité Organizador - The XIII Symposium on Bioinformatics (**JBI 2016**). Valencia, España, mayo 09-13, 2016.
- Miembro del Comité Organizador - The 8th IFIP WG 8.1 working conference on the Practice of Enterprise Modelling (**PoEM 2015**). Valencia, España, noviembre 09-12, 2015.
- Miembro del Comité Organizador - The 25th International Conference on Advanced Information Systems Engineering (**CAiSE 2013**). Valencia, Spain, junio 19-21, 2013.

Premios relevantes de la investigación:

- **Best Poster Award**, 11th International Conference on Research Challenges in Information Science (RCIS 2017), Brighton, UK, Mayo 10-12, 2017. Título: "*GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma*".
- **Best Poster Award**, 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017), Porto, Portugal, Abril 28-30, 2017. Título: "*Software Engineering and Genomics: The Two Sides of the Same Coin?*".
- **Best Oral Presentation Award** (*selected by the Public*), III Meeting of PhD Students at UPV, Valencia, España, Junio 30, 2016. Título: "*Haplotypes and Statistical Models: Integrating to the Conceptual Schema of the Human Genome (CSHG)*".
- **Best Poster Award**, 7th. International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2016), Roma, Italia, Febrero, 21-23, 2016. Título: "*Use of GeIS for Early Diagnosis of Alcohol Sensitivity*".
- **Best Poster Award** (*selected by the Award Committee*), II Meeting of PhD Students at UPV, Valencia, España, Junio 25, 2015. Título: "*Haplotypes Treatment: An Extension to the Conceptual Scheme of Human Genome to Develop GeIS*".

7.3 Trabajo futuro

El trabajo desarrollado en esta Tesis Doctoral demuestra como la aplicación de modelado conceptual en el dominio genómico contribuye a la creación de *Sistemas de Información Genómicos* eficientes, los cuales repercuten de manera directa en la *Medicina Personalizada* o de precisión.

A partir del trabajo del realizado en la presente tesis, surgen distintas líneas de trabajo futuro para tratar:

- **Continuar el desarrollo del Modelo Conceptual del Genoma Humano (MCGH)**: esta tarea consistiría en el estudio y análisis de la versión más reciente, con el objetivo de buscar rutas de crecimiento o extensión del modelo. Un ejemplo, sería la integración de toda la información *clínica* utilizada por los expertos, pues actualmente en el modelo sólo se contempla la parte genómica. Otro punto importante en este objetivo se basa en el análisis del modelo para evaluar su capacidad de adaptabilidad en otras especies -*distinta a la humana*- (por ejemplo[43], el ratón o pez cebra).
    - Incluir en el modelo el conocimiento existente sobre "*Haplogrupos*": Un haplogrupo es un grupo grande de haplotipos, es decir, una combinación de alelos de diferentes loci de un cromosoma que son transmitidos juntos. En genética humana, los haplogrupos más comúnmente estudiados son los haplogrupos del cromosoma Y (ADN-Y) y los haplogrupos del ADN mitocondrial (ADNmt), que pueden ser usados para definir poblaciones genéticas [193].

- **Desarrollo de mecanismos de verificación de la información en fuente (origen)**: tras la obtención de los datos de los distintos repositorios genómicos, y su posterior carga en la base de datos del genoma humano (HGDB) resulta interesante y relevante contar con mecanismos que permitan la actualización automática de los datos conforme a los nuevos datos reportados en las fuentes.

[43] Según Ensembl (*https://www.ensembl.org/index.html*), las tres especies más estudiadas (favoritas) son: humana, ratón y pez cebra.

- **Desarrollo de VarSeach 2.0**: como se ha comentado en este trabajo, la versión inicial de *VarSearch* funciona como un prototipo que proyecta un segmento de las vistas que componen el MCGH. Por lo que, esta tarea está orientada al desarrollo e implementación de las vistas restantes del modelo conceptual (MCGH), como, por ejemplo, la aplicación de los datos haplotípicos presentados en esta tesis para el diagnóstico genómico.

  Actualmente, se ha planteado el desarrollo de un trabajo de investigación (*tesis de máster*) orientado a satisfacer este objetivo. En el trabajo de fin de máster[44] de Alberto García Simón se presentará la versión 2.0 de *VarSearch*. Esta nueva versión incluiría la aplicación de mejoras (rendimiento) y la extensión de la herramienta mediante el tratamiento de nuevas vistas del modelo conceptual del genoma humano.

Dentro del grupo genoma del Centro PROS[45] se están desarrollando otros proyectos de investigación enfocados en:

- La aplicación de métricas de calidad para mejorar la gestión de los datos contenidos en la base de datos del genoma humano (HGDB) [187].
- La integración de mecanismos de "*interacción y colaboración*" aplicados al dominio genómico para mejorar la experiencia -*usuario/ordenador*-, en el tema de interfaces de usuarios [194].

Todo el esfuerzo realizado en este dominio, más las colaboraciones con expertos y especialistas de la materia tiene como objetivo fundamental desarrollar *Sistemas de Información Genómicos* (GeIS) apoyados en *Modelos Conceptuales* para ayudar a la toma de decisiones en el entorno bioinformático.

[44] *Máster Universitario en Ingeniería Informática de la UPV*

[45] *http://www.pros.webs.upv.es/*

Referencias Bibliográficas

[1] A. Olivé, *Conceptual Modeling of Information Systems*, 1st ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[2] R. Wieringa, "Design science methodology," in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - ICSE '10*, 2010, vol. 2, p. 493.

[3] R. Wieringa, "Design science as nested problem solving," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 2009, p. 1.

[4] R. Wieringa, "Introduction to design science methodology," *REFSQ Dr. Symp.*, pp. 1–17, 2013.

[5] R. Wieringa, "Design science research in information systems and software systems engineering," *ICWE 8th*, no. June, 2016.

[6] M. J. Villanueva Del Pozo, "An agile model-driven method for involving end-users in DSL development," Universitat Politècnica de València, Valencia (Spain), 2016.

[7] J. F. Reyes Román, A. León Palacio, and Ó. Pastor López, "Software Engineering and Genomics: The Two Sides of the Same Coin?," in *Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017)*, 2017, no. Enase, pp. 301–307.

[8] Ó. Pastor López, A. León Palacio, J. F. Reyes Román, and J. C. Casamayor, "Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome," in *Conceptual Modeling Perspectives*, J. Cabot, C. Gómez, O. Pastor, M. R. Sancho, and E. Teniente, Eds. Springer International Publishing AG, 2017, pp. 25–40.

[9] J. M. Martínez Rodríguez, "Secuenciación de genomas," *Arbor*, vol. 698, no. Febrero, pp. 285–310, 2004.

[10] Instituto Nacional de Investigación del Genoma Humano (NHGRI), "Terminación del Proyecto Genoma Humano: Preguntas más frecuentes," 2011. [Online]. Available: https://www.genome.gov/11510905/preguntas-maacutes-frecuentes/. [Accessed: 22-May-2017].

[11] M. Gómez Vera, "Unidad 3: Genética Molecular," 2016. [Online]. Available: https://www.slideshare.net/martabiogeo/gentica-molecular-69313111. [Accessed: 03-Nov-2017].

[12] Z. D. Stephens *et al.*, "Big Data: Astronomical or Genomical?," *PLOS Biol.*, vol. 13, no. 7, p. e1002195, Jul. 2015.

[13] J. Escribano, "Estructura del genoma humano : perspectivas en biomedicina," pp. 1–7, 2008.

[14] Genética General: Genoma -Apuntes-, "El mapa del genoma humano," 2012. [Online]. Available: http://www.vi.cl/foro/topic/1389410-genetica-general-genoma-

apuntes/. [Accessed: 24-May-2017].

[15] C. Yusá, "La cronología del descubrimiento del Genoma Humano:," 2010. [Online]. Available: https://www.timetoast.com/timelines/la-cronologia-del-descubrimiento-del-genoma-humano. [Accessed: 20-May-2017].

[16] El País, "LA REVOLUCIÓN GENÉTICA: El desciframiento del genoma humano abre una nueva era en la lucha contra las enfermedades," 2000. [Online]. Available: http://elpais.com/diario/2000/06/27/sociedad/962056801_850215.html. [Accessed: 24-May-2017].

[17] A. D. Baxevanis and B. F. F. Ouellette, *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins (3Ed).*, vol. 34, no. 6. 2006.

[18] M. J. Villanueva, A. R. Guzman, F. Valverde, and A. M. Levin, "Diagen: A model-based bioinformatic tool for genetic analysis," in *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)*, 2012, pp. 1–2.

[19] R. Santamaría, "Secuenciación," *the Science Creative Quarterly*, 2003. [Online]. Available: http://www.scq.ubc.ca/genome-projects-uncovering-the-blueprints-of-biology/.

[20] P. Dawyndt, T. Dedeurwaerdere, and J. Swings, "Exploring and exploiting microbiological commons: contributions of bioinformatics and intellectual property rights in sharing biological information," *Int. Soc. Sci. J*, vol. 188, p. 258, 2004.

[21] Genetic Alliance, *Cómo entender la genética: una guía para pacientes y profesionales médicos en la región de Nueva York y el Atlántico Medio*. 2009.

[22] Instituto Nacional de Investigación del Genoma Humano (NHGRI), "Preguntas frecuentes sobre las pruebas genéticas," 2015. [Online]. Available: https://www.genome.gov/27562619/. [Accessed: 24-May-2017].

[23] T. Kushnick, *Genetics in Medicine*, vol. 267, no. 15. 1992.

[24] T. D. Gelehrter, F. S. Collins, and D. Ginsburg, *Principles of Medical Genetics*, 2nd ed. Lippincott Williams & Wilkins, 1998.

[25] ary B. Mahowald, T. Aspinwall JD, V. A. McKusick MD, and A. Scheuerle MD, *Genetics in the Clinic: Clinical, Ethical, and Social Implications for Primary Care, 1e 1st Edition*. Mosby, 2001.

[26] V. del Castillo Ruiz, R. D. Uranga Hernández, and G. Zafra de la Rosa, *Genética Clínica*. 2013.

[27] W. H. Belloso and M. A. Redal, "LA FARMACOGENOMICA Y EL CAMINO HACIA LA MEDICINA PERSONALIZADA (Artículo Especial)," *Rev. Med. Chil.*, vol. 139, no. 2, pp. 267–273, Feb. 2011.

[28] Boletín Oficial del Estado, "Ley Orgánica 15/1999, de 13 de

diciembre, de Protección de Datos de Carácter Personal.," *Boletín Of. del Estado*, vol. 289, pp. 1–21, 2011.

[29] Boletín Oficial del Estado, "Ley 14/2007, de 3 de julio, de Investigación biomédica.," pp. 1–24, 2012.

[30] I. López-Abadía, "PERIODOS DE CONSERVACIÓN DE DATOS PERSONALES EN INVESTIGACIÓN BIOMÉDICA," 2011.

[31] Boletín Oficial del Estado, "Ley 14/1986, de 25 de abril, General de Sanidad.," *Boe*, no. 102, p. 15207 a 15224, 1986.

[32] BioCore, "Secuenciación de EXOMA Completo," 2017. [Online]. Available: http://www.biocorelabs.com/servicios/enfermedades-geneticas/secuenciacion-de-exoma/. [Accessed: 25-May-2017].

[33] M. Eugenia, B. Nieva, J. Navajas, and J. Cruz, "Artículo especial Diagnóstico prenatal y array-CGH II: gestaciones de bajo riesgo," *Diagnóstico Prenat.*, vol. 23, no. 2, pp. 49–55, 2012.

[34] M. Rodríguez-Rivera, "Técnica de CGH- array," Barcelona, 2017.

[35] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends Genet.*, vol. 30, no. 9, pp. 418–426, Sep. 2014.

[36] B. Rodríguez-Santiago and L. Armengol, "Tecnologías de secuenciación de nueva generación en diagnóstico genético pre- y postnatal," *Diagnóstico Prenat.*, vol. 23, no. 2, pp. 56–66, Apr. 2012.

[37] A. Aguilar, "Medicina Personalizada, Medicina De Precisión, ¿Cuán Lejos Estamos De La Perfección?," *Carcinos*, p. 2, 2015.

[38] Instituto Nacional del Cáncer, "Medicina de precisión en el tratamiento del cáncer," 2015. [Online]. Available: https://www.cancer.gov/espanol/cancer/tratamiento/tip%0Aos/medicina-de-precision. [Accessed: 01-Jan-2017].

[39] N. Jiménez, "Una medicina nueva, más inteligente y menos invasiva," 2014. [Online]. Available: http://www.lifescienceslab.com. [Accessed: 01-Jan-2017].

[40] J. G. Paez *et al.*, "EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy," *Science (80-.).*, vol. 304, no. 5676, pp. 1497–1500, Jun. 2004.

[41] S. J. Aronson and H. L. Rehm, "Building the foundation for genomics in precision medicine," *Nature*, vol. 526, no. 7573, pp. 336–342, 2015.

[42] M. Posada and I. Abaitua, "Enfermedades raras . Concepto, epidemiología y situación actual en España," vol. 31, 2008.

[43] E. R. Mardis, "The $1,000 genome, the $100,000 analysis?," *Genome Med.*, vol. 2, no. 11, p. 84, 2010.

[44] D. Aguilera, C. Gómez, and A. Olivé, "Enforcement of Conceptual Schema Quality Issues in Current Integrated Development

Environments," no. June, 2013, pp. 626–640.

[45] N. W. Paton *et al.*, "Conceptual modelling of genomic information," vol. 16, no. 6, pp. 548–557, 2000.

[46] N. W. Paton, E. Bornberg-bauer, and N. W. Paton, "Conceptual data modelling for bioinformatics," vol. 3, no. 2, pp. 166–180, 2002.

[47] S. Ram and W. Wei, "Modeling the Semantics of 3D Protein Structures," in *Genome*, 2004, pp. 696–708.

[48] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, 2000.

[49] Gene Ontology Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Res.*, vol. 32, no. 90001, p. 258D–261, 2004.

[50] O. Pastor, S. España, J. I. Panach, and N. Aquino, "Model-Driven Development," *Informatik-Spektrum*, vol. 31, no. 5, pp. 394–407, Oct. 2008.

[51] K. Garwood *et al.*, "Model-driven user interfaces for bioinformatics data resources: regenerating the wheel as an alternative to reinventing it," *BMC Bioinformatics*, vol. 7, no. 1, p. 532, 2006.

[52] M. Rouse, "A data lake is a large object based storage repository that holds data in its native format until it is needed," 2014.

[53] J. D. Breul, *Cyber Society, Big Data, and Evaluation : Comparative Policy Evaluation*. Transaction Publishers, 2017.

[54] D. J. Rigden, X. M. Fernández-Suárez, and M. Y. Galperin, "The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1–D6, Jan. 2016.

[55] E. Alexov, "Navigating through Genomics Data to Deliver Testable Predictions," *Hum. Mutat.*, vol. 36, no. 2, pp. v–v, Feb. 2015.

[56] A. Auton *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Sep. 2015.

[57] C. Rodriguez and M. Aluztisa, "El Proyecto 1000 Genomas," *Medicina*, 2016. [Online]. Available: http://medmol.es/noticias/279/. [Accessed: 05-Jun-2017].

[58] R. M. Durbin *et al.*, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[59] P. H. Sudmant *et al.*, "An integrated map of structural variation in 2,504 human genomes," *Nature*, vol. 526, no. 7571, pp. 75–81, Sep. 2015.

[60] M. V Osier, K. H. Cheung, J. R. Kidd, A. J. Pakstis, P. L. Miller, and K. K. Kidd, "ALFRED: an allele frequency database for diverse populations and DNA polymorphisms--an update.,"

Nucleic Acids Res., vol. 29, no. 1, pp. 317–319, 2001.

[61] K. K. Kidd, "The ALlele FREquency Database (ALFRED)," *NSF Grant*, p. 2, 2016.

[62] H. Rajeevan, "ALFRED: the ALelle FREquency Database. Update," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 270–271, Jan. 2003.

[63] C. Szabo, A. Masiello, J. F. Ryan, and L. C. Brody, "The Breast Cancer Information Core: Database design, structure, and scope," *Hum. Mutat.*, vol. 16, no. 2, pp. 123–131, 2000.

[64] S. F. Saccone, J. Quan, and P. L. Jones, "BioQ: tracing experimental origins in public genomic databases using a novel data provenance model," *Bioinformatics*, vol. 28, no. 8, pp. 1189–1191, Apr. 2012.

[65] M. J. Landrum *et al.*, "ClinVar: Public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 980–985, 2014.

[66] M. J. Landrum *et al.*, "ClinVar: public archive of interpretations of clinically relevant variants," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D862–D868, Jan. 2016.

[67] S. Bamford *et al.*, "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website," *Br. J. Cancer*, vol. 2, pp. 355–358, Jun. 2004.

[68] S. A. Forbes *et al.*, "COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer," *Nucleic Acids Res.*, vol. 38, no. suppl_1, pp. D652–D657, Jan. 2010.

[69] S. A. Forbes *et al.*, "COSMIC: exploring the world's knowledge of somatic mutations in human cancer," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D805–D811, Jan. 2015.

[70] K. A. Tryka *et al.*, "NCBI's database of genotypes and phenotypes: DbGaP," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 975–979, 2014.

[71] S. T. Sherry, "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, Jan. 2001.

[72] L. Phan *et al.*, "dbSNP and dbVar : NCBI Databases of Simple and Structural Variations," p. 20894, 2015.

[73] E. Masood, ". . . as consortium plans free SNP map of human genome," *Nature*, vol. 398, no. 6728, pp. 545–545, Apr. 1999.

[74] A. J. Brookes, "HGBASE: a database of SNPs and other variations in and around human genes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 356–360, Jan. 2000.

[75] K. Higasa, K. Miyatake, Y. Kukita, T. Tahira, and K. Hayashi, "D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples," *Nucleic Acids*

Res., vol. 35, no. Database, pp. D685–D689, Jan. 2007.

[76] T. Tahira *et al.*, "A definitive haplotype map of structural variations determined by microarray analysis of duplicated haploid genomes," *Genomics Data*, vol. 2, pp. 55–59, Dec. 2014.

[77] J. Pinero *et al.*, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, p. bav028-bav028, Apr. 2015.

[78] T. Hubbard *et al.*, "The Ensembl genome database project," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 38–41, 2002.

[79] F. Cunningham *et al.*, "Ensembl 2015," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D662–D669, Jan. 2015.

[80] W. McLaren *et al.*, "The Ensembl Variant Effect Predictor," *Genome Biol.*, vol. 17, no. 1, p. 122, 2016.

[81] R. Oliva Virgili and J.-M. Vidal-Taboada, *Genoma humano. Nuevos avances en investigación, diagnóstico y tratamiento*. 2006.

[82] Instituto Nacional de Investigación del Genoma Humano (NHGRI), "Acerca del Proyecto Internacional HapMap," 2015. [Online]. Available: https://www.genome.gov/27562906/acerca-del-proyecto-internacional-hapmap/. [Accessed: 30-May-2017].

[83] R. A. Gibbs *et al.*, "The international HapMap project," *Nature*, vol. 426, no. 6968, pp. 789–796, 2003.

[84] International HapMap Consortium, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, p. 1299–320 ST–A haplotype map of the human genome, 2005.

[85] Coriell Institute for Medical Research, "Noticias HapMap," *Coriell Inst. Med. Res.*, vol. 3, 2007.

[86] HGMD, "Introducción a HGMD®," 2017. [Online]. Available: http://www.hgmd.cf.ac.uk/ac/introduction.php?lang=spanish. [Accessed: 06-Jun-2017].

[87] P. D. Stenson, M. Mort, E. V. Ball, K. Shaw, A. D. Phillips, and D. N. Cooper, "The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine," *Hum. Genet.*, vol. 133, no. 1, pp. 1–9, Jan. 2014.

[88] P. D. Stenson *et al.*, "The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies," *Hum. Genet.*, vol. 136, no. 6, pp. 665–677, Jun. 2017.

[89] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.*, vol. 27, no. 1, pp. 29–34, Jan. 1999.

[90] I. F. A. C. Fokkema, P. E. M. Taschner, G. C. P. Schaafsma, J. Celli, J. F. J. Laros, and J. T. den Dunnen, "LOVD v.2.0: The

next generation in gene variant databases," *Hum. Mutat.*, vol. 32, no. 5, pp. 557–563, 2011.

[91] I. F. A. C. Fokkema and D. Asscheman, "LOVD 3.0 user manual (Build 3.0-17)," 2016.

[92] A. Hamosh, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D514–D517, Dec. 2004.

[93] V. A. McKusick, "Mendelian Inheritance in Man and Its Online Version, OMIM®," *Am. J. Hum. Genet.*, vol. 80, no. 4, pp. 588–604, Apr. 2007.

[94] J. Amberger, C. Bocchini, and A. Hamosh, "A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®)," *Hum. Mutat.*, vol. 32, no. 5, pp. 564–567, May 2011.

[95] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D789–D798, Jan. 2015.

[96] A. Fabregat *et al.*, "The reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D481–D487, Jan. 2016.

[97] D. Croft *et al.*, "The Reactome pathway knowledgebase," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D472–D477, Jan. 2014.

[98] S. Jupe, A. Fabregat, and H. Hermjakob, "Expression Data Analysis with Reactome," in *Current Protocols in Bioinformatics*, vol. 49, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015, p. 8.20.1-8.20.9.

[99] G. Rodríguez Tarduchy, *¿Hablamos de gen…o…mas?* Editorial Hélice, 2007.

[100] M. Cariaso and G. Lennon, "SNPedia: a wiki supporting personal genome annotation, interpretation and analysis," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D1308–D1312, Jan. 2012.

[101] B. M. Good, E. L. Clarke, S. Loguercio, and A. I. Su, "Linking genes to diseases with a SNPedia-Gene Wiki mashup," *J. Biomed. Semantics*, vol. 3, no. Suppl 1, p. S6, 2012.

[102] P. A. Fujita *et al.*, "The UCSC Genome Browser database: update 2011," *Nucleic Acids Res.*, vol. 39, no. Database, pp. D876–D882, Jan. 2011.

[103] L. R. Meyer *et al.*, "The UCSC Genome Browser database: extensions and updates 2013," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D64–D69, Jan. 2013.

[104] D. A. Benson *et al.*, "GenBank," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D36–D42, Jan. 2013.

[105] M. L. Speir *et al.*, "The UCSC Genome Browser database: 2016 update," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D717–D725, Jan.

2016.

[106] C. Béroud, D. Hamroun, G. Collod-Béroud, C. Boileau, T. Soussi, and M. Claustres, "UMD (Universal Mutation Database): 2005 Update," *Hum. Mutat.*, vol. 26, no. 3, pp. 184–191, 2005.

[107] C. Béroud, G. Collod-Béroud, C. Boileau, T. Soussi, and C. Junien, "UMD (Universal Mutation Database): A generic software to build and analyze locus-specific databases," *Hum. Mutat.*, vol. 15, no. 1, pp. 86–94, 2000.

[108] A. Bateman *et al.*, "UniProt: a hub for protein information," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D204–D212, Jan. 2015.

[109] S. Willuweit and L. Roewer, "Y chromosome haplotype reference database (YHRD): Update," *Forensic Sci. Int. Genet.*, vol. 1, no. 2, pp. 83–87, 2007.

[110] S. Willuweit and L. Roewer, "The new Y Chromosome Haplotype Reference Database," *Forensic Sci. Int. Genet.*, vol. 15, pp. 43–48, Mar. 2015.

[111] L. Roewer, "The Y-Chromosome Haplotype Reference Database (YHRD)—Publicly Available Reference and Research Datasets for the Forensic Interpretation of Y-Chromosome STR Profiles," *Handb. Forensic Genet. Biodivers. Hered. Civ. Crim. Investig.*, pp. 231–248, 2017.

[112] J. F. Reyes Román, Ó. Pastor, J. C. Casamayor, and F. Valverde, "Applying Conceptual Modeling to Better Understand the Human Genome," in *ER 2016: Conceptual Modeling*, vol. 9974, Gifu, Japan: Springer International Publishing, 2016, pp. 404–412.

[113] O. Pastor, A. M. Levin, M. Celma, J. C. Casamayor, A. Virrueta, and L. E. Eraso, "Model-Based Engineering Applied to the Interpretation of the Human Genome," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6520, 2011, pp. 306–330.

[114] J. T. den Dunnen *et al.*, "HGVS Recommendations for the Description of Sequence Variants: 2016 Update," *Hum. Mutat.*, vol. 37, no. 6, pp. 564–569, 2016.

[115] Nature, "Open reading frames," 2017. [Online]. Available: https://www.nature.com/subjects/open-reading-frames. [Accessed: 18-Aug-2017].

[116] A. Zeron, "Biotipos, fenotipos y genotipos. ¿Qué biotipo tenemos?," *Rev. Mex. Periodontol.*, vol. 2, no. 1, pp. 22–33, 2011.

[117] S. Jablonski, "Syndrome: Le Mot de Jour," *Am. J. Med. Genet.*, vol. 39, no. 3, pp. 342–346, Jun. 1991.

[118] Genetics Home Reference, "What are single nucleotide polymorphisms (SNPs)?," 2017. [Online]. Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp. [Accessed:

13-Jun-2017].

[119] K. del C.-M. Ángel Lugo-Trampe, "Medicina Universitaria," vol. 11, no. 44, pp. 187–192, 2009.

[120] J. F. Reyes Román, Ó. Pastor, F. Valverde, and D. Roldán M., "How to deal with Haplotype data : An Extension to the Conceptual Schema of the Human Genome," vol. 19, no. 3, pp. 1–21, 2016.

[121] J. F. Reyes Román, Ó. Pastor, F. Valverde, and D. Roldán M., "Including haplotypes treatment in a Genomic Information Systems Management," in *Ibero-American Conference on Software Engineering*, 2016, pp. 11–24.

[122] J. F. Reyes Román and Ó. Pastor, "Use of GeIS for Early Diagnosis of Alcohol Sensitivity," in *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 3: BIOINFORMATICS*, 2016, vol. 3, no. Biostec, pp. 284–289.

[123] A. M. Sánchez-Pérez, "Sensibilidad al Alcohol y la Predisposición a beber," *Ciencia al Día Internacional*, 2000. [Online]. Available: http://www.ciencia.cl/CienciaAlDia/volumen3/numero2/articulos/articulo3.html.

[124] D. M. Dick and T. Foroud, "Candidate Genes for Alcohol Dependence: A Review of Genetic Evidence From Human Studies," *Alcohol. Clin. Exp. Res.*, vol. 27, no. 5, pp. 868–879, 2003.

[125] E. García Gutiérrez, G. Lima Mompó, L. Aldana Vilas, P. Casanova Carrillo, and V. Feliciano Álvarez, "Alcoholismo y sociedad, tendencias actuales," *Rev. Cuba. Med. Mil.*, vol. 33, no. 3, 2014.

[126] The Scripps Research Institute, "The Effects of Alcohol on the Brain," 2016. [Online]. Available: http://www.scripps.edu/newsandviews/e_20020225/koob2.html. [Accessed: 01-Jan-2016].

[127] J. F. Reyes Román, Ó. Pastor, and M. R. Fernández Alcalá, "Integración de haplotipos al modelo conceptual del genoma humano utilizando la metodología sile," Universitat Politecnica de Valencia, 2014.

[128] PubMed, "PubMed," *Webpage*, 2017. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed. [Accessed: 01-Jan-2017].

[129] M. Soyka, U. W. Preuss, V. Hesselbrock, P. Zill, G. Koller, and B. Bondy, "GABA-A2 receptor subunit gene (GABRA2) polymorphisms and risk for alcohol dependence," *J. Psychiatr. Res.*, vol. 42, no. 3, pp. 184–191, 2008.

[130] H. J. Edenberg *et al.*, "Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol

dependence and with brain oscillations.," *Am. J. Hum. Genet.*, vol. 74, no. 4, pp. 705–714, 2004.

[131] G. J. Lydall *et al.*, "Genetic association study of GABRA2 single nucleotide polymorphisms and electroencephalography in alcohol dependence," *Neurosci. Lett.*, vol. 500, no. 3, pp. 162–166, 2011.

[132] P. L. Fernández, J. M. Ladero Quesada, J. C. Leza Cerro, and I. Lizasoain H., *Drogodependencias: Farmacología - Patología – Psicología – Legislación*, 3rd ed. Ed. Médica Panamericana, 2009.

[133] B. Campos, O. Díez, C. Álvarez, L. Palma, J. Balmaña, and P. Carvallo, "Análisis del haplotipo en portadores de la mutación 6857delAA en el gen BRCA2 en 4 familias con cáncer de mama u ovario hereditario," *Med. Clin. (Barc).*, vol. 123, no. 14, pp. 543–545, 2004.

[134] A. S. Álvarez, J. T. Márquez, F. R. Vargas, and M. R. Romero, "Asociación del cáncer de mama con los polimorfismos T-66G y G-156GG del gen SPP1 y las concentraciones séricas de osteopontina," *Ginecol. Obstet. Mex.*, vol. 80, no. 1, pp. 22–29, 2012.

[135] D. Fonseca, C. Silva, H. Mateus, and C. M. Restrepo, "Identificación de deleciones en portadoras de distrofia muscular de Duchenne. Deletions identification in female carriers of Duchenne ' s muscular distrophy," pp. 63–67, 2008.

[136] S. B. Gabriel *et al.*, "The structure of haplotype blocks in the human genome," *Science (80-.).*, vol. 296, no. 2002, pp. 2225–2229, 2011.

[137] P. Delves, S. Martin, D. Burton, and I. Roitt, *Inmunología: Fundamentos*, 12th ed. 2014.

[138] NCBI, "dbSNP ER Schema," 2015. [Online]. Available: ftp://ftp.ncbi.nih.gov/snp/database/b124/mssql/schema/erd_dbSNP.pdf. [Accessed: 01-Jan-2015].

[139] Ensembl, "Ensembl: Features_Analyses_Core Schema," 2016. [Online]. Available: http://www.ensembl.org/info/docs/api/core/features_analyses_core.pdf.

[140] UCSC, "UCSC Genome Bioinformatics: Schema for Haplotypes - GRCh38 Haplotype to Reference Sequence Mapping Correspondence," 2016. [Online]. Available: http://ucscbrowser.genap.ca/cgi-bin/hgTables?db=hg38&hgta_group=map&hgta_track=altLocations&hgta_table=altLocations&hgta_doSchema=describe+table+schema.

[141] W. J. Kent *et al.*, "Exploring relationships and mining data with the UCSC Gene Sorter," *Genome Res.*, vol. 15, no. 5, pp. 737–741, 2005.

[142] UCSC, "UCSC Genome Bioinformatics: Haplotypes representation," 2015. [Online]. Available: https://genome.ucsc.edu/goldenPath/help/haplotypes.html.

[143] A. M. Martínez Ferrandis, "Wiki-Genome:A model-driven genome data management environment," in *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)*, 2011.

[144] C. J. Date and S. L. M. Ruiz Faudón, *INTRODUCCION A LOS SISTEMAS DE BASES DE DATOS (7ª ED.)*, 7th ed. S.A. ALHAMBRA MEXICANA, 2001.

[145] The Sequence Ontology, "Haplotype (Current_SVN)," 2016. [Online]. Available: http://sequenceontology.org/browser/current_svn/term/SO:0001024. [Accessed: 01-Jan-2016].

[146] L. Muñoz, J. N. Mazón, and J. Trujillo, "ETL Process Modeling Conceptual for Data Warehouses : A Systematic Mapping Study," *America (NY).*, vol. 9, no. 3, pp. 358–363, 2011.

[147] ETL-Tools.Info, "Proceso ETL," 2016. [Online]. Available: http://etl-tools.info/es/bi/proceso_etl.htm. [Accessed: 01-Jan-2016].

[148] D. Fallin and N. J. Schork, "Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data.," *Am. J. Hum. Genet.*, vol. 67, no. 4, pp. 947–959, 2000.

[149] M. Stephens, N. J. Smith, and P. Donnelly, "A new statistical method for haplotype reconstruction from population data.," *Am. J. Hum. Genet.*, vol. 68, no. 4, pp. 978–989, 2001.

[150] D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland, "Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous," *Am. J. Hum. Genet*, vol. 70, pp. 425–434, 2002.

[151] U. de la R. U. UVIGEN, "Genética de Poblaciones," 2003. [Online]. Available: http://uvigen.fcien.edu.uy/utem/Popgen/popintro.html.

[152] A. Barbadilla, "Genética de Poblaciones," *Departamento de Genética y Microbiología, Universidad Autónoma de Barcelona*, 2009. [Online]. Available: http://bioinformatica.uab.es/divulgacio/genpob.html. [Accessed: 01-Jan-2017].

[153] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: Analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.

[154] Atlas of Genetics and Cytogenetics in Oncology and Haematology, *Modelo de Hardy-Weinberg*. Atlas of Genetics and Cytogenetics in Oncology and Haematology, 2015.

[155] T. Nagylaki, "Fixation indices in subdivided populations," *Genetics*, vol. 148, no. 3, pp. 1325–1332, 1998.

[156] M. Slatkin, "Linkage disequilibrium--understanding the evolutionary past and mapping the medical future.," *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–85, 2008.

[157] S. Teoh, Y. Yang, and Y. Zhang, "R-square and market efficiency," *Available SSRN 926948*, 2009.

[158] A. Carvajal-Rodríguez, "Cálculo del valor de recombinación que maximiza el LOD Score," no. Morton, pp. 1–2, 1955.

[159] Sanger Institute, "HapMap 3," 2016. [Online]. Available: http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html. [Accessed: 10-May-2017].

[160] D. E. Reich *et al.*, "Linkage disequilibrium in the human genome," *Nature*, vol. 411, no. 6834, pp. 199–204, 2001.

[161] J. J. Johnston and L. G. Biesecker, "Databases of genomic variation and phenotypes: Existing resources and future needs," *Hum. Mol. Genet.*, vol. 22, no. R1, 2013.

[162] A. J. Brookes and P. N. Robinson, "Human genotype–phenotype databases: aims, challenges and opportunities," *Nature*, 2015.

[163] D. Roldán M., Ó. Pastor, and J. F. Reyes Román, "E-Genomic Framework for delivering genomic services. An application to JABAWS," in *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*, 2015.

[164] J. F. Reyes Román, C. E. Iñiguez-Jarrín, and Ó. Pastor, "GenesLove.Me : A Model-based Web-application for Direct-to-consumer Genetic Tests," in *Proceedings of the 12th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2017)*, 2017, no. Enase, pp. 133–143.

[165] V. Burriel Coll, J. F. Reyes Román, A. Heredia Casanoves, C. E. Iñiguez-Jarrín, and A. León Palacio, "GeIS based on Conceptual Models for the Risk Assessment of Neuroblastoma," in *2017 IEEE Eleventh International Conference on Research Challenges in Information Science (RCIS)*, 2017, pp. 451–452.

[166] R. Aracil López and Ó. Pastor, "Diseño e Implementación de un Sistema de Información Clínico y Genómico para la Patología de Dupuytren," Universitat Politecnica de Valencia, 2014.

[167] A. Santana Donato, Ó. Pastor, and M. R. Fernández Alcalá, "Modelado e implementacion de las probabilidades de riesgo al modelo conceptual del genoma humano," Universitat Politecnica de Valencia, 2013.

[168] J. Guerola Martínez, Ó. Pastor, and M. R. Fernández Alcalá, "Sistema de información genómico: Extensión del módulo de carga e integración de información genómica," Universitat Politecnica de Valencia, 2013.

[169] J. Muñoz, M. Llacer, and B. Bonet, "Configuring ATL transformations in MOSKitt," in *CEUR Workshop Proceedings*, 2010, vol. 711, pp. 50–59.

[170] NCBI, "NCBI main site," 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/.

[171] National Cancer Institute, "National Cancer Institute (main site)," 2017. [Online]. Available: https://www.cancer.gov/types/breast. [Accessed: 01-Jan-2017].

[172] H. Zhou, D. Yang, and Y. Xu, "An ETL strategy for real-time data warehouse," *Adv. Intell. Soft Comput.*, vol. 124, pp. 329–336, 2011.

[173] M. Van Der Kroon, "Conceptual Modeling Applied to Genomics: Challenges Faced in Data Loading," no. september, 2012.

[174] M. Navarrete Hidalgo, Ó. Pastor, and J. F. Reyes Román, "Diseño e Implementación de un Sistema de Información Genómico para el Diagnóstico de la Catarata Congénita utilizando la Metodología SILE," Universitat Politècnica de València, 2017.

[175] GENETAQ, "BIOINFORMATICA PARA NO INICIADOS: CAPITULO I," 2015. [Online]. Available: http://genetaq.com/es/blog/bioinformatica-para-no-iniciados-capitulo-i. [Accessed: 30-Jun-2017].

[176] Q. F. Info, "The Variant Call Format (VCF) Version 4.2 Specification," pp. 1–28, 2016.

[177] P. Danecek *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.

[178] J. M. Claverie and C. Notredame, *Bioinformatics for dummies*. John Wiley & Sons, 2011.

[179] D. Roldan Martinez, Ó. Pastor, and M. R. Fernández Alcalá, "An integration architecture framework for e- genomlcs services," in *2014 IEEE 8th International Conference on Research Challenges in Information Science (RCIS)*, 2014, pp. 1–7.

[180] J. D. Gauchat, *El gran libro de HTML5, CSS3 y Javascript*. Marcombo, 2012.

[181] Navicat, "Navicat Enterprise tutorial," 2017. [Online]. Available: https://www.navicat.com/es/store/navicat-for-mysql. [Accessed: 01-Jan-2017].

[182] F. Valverde and O. Pastor, "Dealing with REST Services in Model-driven Web Engineering Methods," in *V Jornadas Científico-Técnicas en Servicios Web y SOA, JSWEB*, 2009, no. October, pp. 243–250.

[183] F. Haupt, D. Karastoyanova, F. Leymann, and B. Schroth, "A model-driven approach for REST compliant services," in *2014 IEEE International Conference on Web Services (ICWS)*, 2014.

[184] M. S. Silva, *JQuery: a biblioteca do programador JavaScript*. São

Paulo: Novatec, 2008.

[185] P. Cingolani *et al.*, "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3," *Fly (Austin).*, vol. 6, no. 2, pp. 80–92, 2012.

[186] P. Cingolani *et al.*, "Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift," *Front. Genet.*, vol. 3, no. MAR, pp. 1–9, 2012.

[187] A. León Palacio, J. F. Reyes Román, V. Burriel Coll, and F. Valverde, "Data Quality Problems When Integrating Genomic Information," in *ER 2016 Workshops*, 2016, vol. 7518, pp. 173–182.

[188] K. Wang, M. Li, and H. Hakonarson, "ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data," *Nucleic Acids Res.*, vol. 38, no. 16, pp. 1–7, 2010.

[189] Gembiosoft SME, "Gembiosoft (main site)," 2017. [Online]. Available: http://gembiosoft.com/. [Accessed: 01-Jan-2017].

[190] S. Romeo-Malanda, "Análisis genéticos directos al consumidor: cuestiones éticas y jurídicas.," 2009. [Online]. Available: http://www.institutoroche.es/legalactualidad/85/analisis. [Accessed: 19-Dec-2016].

[191] UNESCO, "Declaración Internacional sobre los Datos Genéticos Humanos," 2004. [Online]. Available: http://unesdoc.unesco.org/images/0013/001361/%0A136112so.pdf. [Accessed: 19-Dec-2016].

[192] Comunidad Autónoma de Galicia, "Ley 3/2001, de 28 de mayo, reguladora del consentimeinto informado y de la historia clínica de los pacientes," *Boletín Of. del Estado*, pp. 23537–23541, 2001.

[193] N. Cosme Boullosa, "¿Y qué fue de Adán y Eva?," pp. 1–25, 2010.

[194] C. E. Iñiguez-Jarrín, A. García Simon, J. F. Reyes Román, and Ó. Pastor, "GenDomus : Interactive and Collaboration Mechanisms for Diagnosing Genetic Diseases," *Proc. 12th Int. Conf. Eval. Nov. Approaches to Softw. Eng. (ENASE 2017)*, no. Enase, pp. 91–102, 2017.

[195] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.

[196] N. Guarino, "Formal Ontology and Information Systems," *Biomed. Environ. Sci.*, vol. 13, no. 1, pp. 37–43, Mar. 2000.

[197] Instituto Nacional del Cáncer, "Definición: Mutación Somática," 2017. [Online]. Available: https://www.cancer.gov/espanol/publicaciones/diccionario?cdrid=46586. [Accessed: 05-Jun-2017].

[198] GDICT, "Significado de Checksum," 2017. [Online]. Available: http://es.gdict.org/definicion.php?palabra=checksum. [Accessed:

20-Oct-2017].

[199] Instituto Nacional de Investigación del Genoma Humano, "Definición de LOCUS, Institutos Nacionales de la Salud," 2016. [Online]. Available: https://www.genome.gov/glossarys/index.cfm?id=116]. [Accessed: 13-Jun-2017].

ANEXOS

Anexo A. Diccionario de Datos:
Base de Datos del Genoma Humano (HGDB)

@

@Bib_Ref_ID	=	{Numérico}	
@Chr_Elem_ID	=	{Numérico}	
@Chr_Exon_ID	=	{Numérico}	
@Chr_Gene_ID	=	{Numérico}	*(NULL)*
@Chr_Transcript_ID	=	{Numérico}	*(NULL)*
@Curator_ID	=	{Numérico}	
@DB_Version_ID	=	{Numérico}	*(NULL)*
@HG_Identifier	=	1{Carácter}9	
@NC_Identifier	=	1{Carácter}15	
@NG_Identifier	=	1{Carácter}15	*(NULL)*
@Nombre	=	1{Carácter}100	
@Phenotype_ID	=	{Numérico}	
@URL	=	1{Carácter}255	
@Variation_ID	=	{Numérico}	

A

Abstract	=	1{Carácter}65535	*(NULL)*
ALN_Quality	=	{Numérico}	*(NULL)*
Associated_Genes	=	1{Carácter}65535	*(NULL)*
Authors	=	1{Carácter}65535	*(NULL)*

B

BIB_REF	=	@Bib_Ref_ID + Title + Abstract + Publication + Authors + Date_Pub	3
Bib_Ref_ID	=	{Numérico}	
BIBLIOGRAPHY_DB	=	@URL + Bib_Ref_ID + Name_DB + Pubmed_ID	3

Biotype	=	1{Carácter}30	*(NULL)*

C

Carácter	=	[0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |9| A-Z | a-z | |. | , | ; | - |?]	
CERTAINTY	=	@Phenotype_ID + Variation_ID + Level_Certainty	2
CHR_ELEM	=	@Chr_Elem_ID + NC_Identifier + Start_Position + End_Position + Strand + Specialization_Type	1
CHROMOSOME	=	@NC_Identifier + Nombre + HG_Identifier + Sequence	1
Clinical_Significance	=	1{Carácter}250	*(NULL)*
Clinically_Important	=	1{Carácter}50	*(NULL)*
Comentario	=	1{Carácter}255	*(NULL)*
Creation_Version	=	1{Carácter}65535	*(NULL)*
CURATOR	=	@Curator_ID + User_Name + Pass	4

D

DATABANK	=	@Nombre + Description + URL	3
DATABANK_VERSION	=	Release + Nombre + Fecha + @DB_Version_ID	3
Date_Pub	=	[dd/mm/aaaa; hh:mm:ss: tt]	*(NULL)*
DB_Variation_ID	=	1{Carácter}250	*(NULL)*
Description	=	1{Carácter}255	*(NULL)*
Description	=	1{Carácter}65535	*(NULL)*
DNA_Sequence	=	1{Carácter}4294967298	*(NULL)*
Double	=	Número en coma flotante de precisión doble. Los valores permitidos van desde -1.7976931348623157E+308 a -2.2250738585072014E-308, 0 y desde 2.2250738585072014E-308 a 1.7976931348623157E+308	

E

ELEMENT_DATABANK	=	@DB_Version_ID + Source_Identifier + @Chr_Elem_ID	3
End_ExonNG	=	{Numérico}	*(NULL)*
End_GeneNG	=	{Numérico}	*(NULL)*
End_Position	=	{Numérico}	*(NULL)*
End_TranscriptNG	=	{Numérico}	*(NULL)*
EndCDS	=	{Numérico}	*(NULL)*
EXON	=	@Chr_Elem_ID + Nombre + ID_Symbol + Start_ExonNG + End_ExonNG	1
EXON_TRANSCRIPT	=	@Chr_Exon_ID + @Chr_Transcript_ID + Nombre	1

F

Fecha	=	[dd/mm/aaaa; hh:mm:ss: tt]	*(NULL)*
Flanking_Left	=	1{Carácter}25	*(NULL)*
Flanking_Right	=	1{Carácter}25	*(NULL)*

G

GC_Percentage	=	{Double}	*(NULL)*
GENE	=	@Chr_Elem_ID + ID_Symbol + ID_HUGO + Official_Name + Description + Biotype + Status + GC_Percentage + Gene_Synonym + Start_GeneNG + End_GeneNG	1
Gene_Synonym	=	1{Carácter}250	*(NULL)*
GENOME	=	@HG_Identifier + GRCH_Identifier	1
GRCH_Identifier	=	1{Carácter}13	

H

HG_Identifier	=	1{Carácter}9	*(NULL)*

I

ID_HUGO	=	1{Carácter}10	*(NULL)*
ID_Symbol	=	1{Carácter}10	*(NULL)*
IMPRECISE	=	@Variation_ID + Description	2
INS_Repetition	=	{Numérico}	*(NULL)*
INS_Sequence	=	1{Carácter}65535	*(NULL)*
Internal_Code	=	1{Carácter}50	*(NULL)*

L

Level_Certainty	=	0{Carácter}1	*(NULL)*

N

Name	=	1{Carácter}100	
Name_DB	=	1{Carácter}255	*(NULL)*
NC_Identifier	=	1{Carácter}15	*(NULL)*
NG_Identifiers	=	1{Carácter}65535	*(NULL)*
NM_Identifier	=	1{Carácter}15	*(NULL)*
Nombre	=	1{Carácter}5	*(NULL)*
Nombre	=	1{Carácter}100	*(NULL)*
Nombre	=	1{Carácter}20	*(NULL)*
Nombre	=	1{Carácter}30	*(NULL)*
Nombre	=	1{Carácter}255	*(NULL)*
NP_Identifier	=	1{Carácter}15	*(NULL)*
NUM_Bases	=	{Numérico}	*(NULL)*
Numérico	=	[0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9] *Almacena números enteros y decimales*	

O

Official_Name	=	1{Carácter}250	*(NULL)*
OMIM	=	1{Carácter}65535	*(NULL)*
Other_Identifiers	=	1{Carácter}65535	*(NULL)*

P

Pass	=	1{Carácter}30	*(NULL)*
PHENOTYPE	=	@Phenotype_ID + Nombre	2
Position	=	{Numérico}	*(NULL)*
PRECISE	=	@Variation_ID + Specialization_Type + INS_Sequence + INS_Repetition + NUM_Bases + Flanking_Right + Flanking_Left + ALN_Quality + Position	2
PRECISE_SEQNG	=	@Variation_ID + @NG_Identifier + Position + Flanking_Left + Flanking_Right	2
Privado	=	{Numérico}	*(NULL)*
Privado	=	{Numérico: [0 | 1]}	*(NULL)*
PROTEIN	=	@Chr_Trancript_ID + @Name + Sequence + Source + NP_Identifier	1
Publication	=	1{Carácter}65535	*(NULL)*
Pubmed_ID	=	{Numérico}	*(NULL)*

R

REF_CHR_ELEM	=	@Chr_Elem_ID + @Bib_Ref_ID	3
REFERENCE_VARIATION	=	@Variation_ID + @Bib_Ref_ID	3
Release	=	1{Carácter}255	*(NULL)*

S

Sequence	=	1{Carácter}4294967298	*(NULL)*
SEQUENCE_NG	=	DNA_Sequence + ID_Symbol + @NG_Identifier	1
Source	=	1{Carácter}100	*(NULL)*
Source_Identifier	=	1{Carácter}20	*(NULL)*
Specialization_Type	=	1{Carácter}25	*(NULL)*
Specialization_Type	=	[ID | DE | IS | IN]	*(NULL)*
Start_ExonNG	=	{Numérico}	*(NULL)*
Start_GeneNG	=	{Numérico}	*(NULL)*
Start_Position	=	{Numérico}	*(NULL)*
Start_TranscriptNG	=	{Numérico}	*(NULL)*
StartCDS	=	{Numérico}	*(NULL)*
Status	=	1{Carácter}30	*(NULL)*
Strand	=	[P | M]	*(NULL)*

T

Title	=	1{Carácter}65535	*(NULL)*
TRANSCRIPT	=	@Chr_Elem_ID + Biotype + StartCDS + EndCDS + NM_Identifier + @NG_Identifier + Start_TranscriptNG + End_TranscriptNG	1

U

URL	=	1{Carácter}100	*(NULL)*
User_Name	=	1{Carácter}20	*(NULL)*

V

VALIDATION	=	@Variation_ID + @Curator_ID + Comentario + Privado +	4

		Validation_Date + Internal_Code + Clinical_Significance	
Validation_Date	=	[dd/mm/aaaa; hh:mm:ss: tt]	*(NULL)*
VARIATION	=	@Variation_ID + @Chr_Gene_ID + @DB_Version_ID + Description + DB_Variation_ID + Clinically_Important + Privado + NC_Identifier + NG_Identifiers + Other_Identifiers + Associated_Genes + OMIM + Creation_Version	2
Variation_ID	=	{Numérico}	

1. Estructural (CORE); **2.** Variaciones; **3.** Fuentes de Datos y Bibliografía; **4.** Usuarios y Validaciones

Anexo B. Glosario[46]

A

ADN

ADN es el nombre químico de la molécula que contiene la información genética en todos los seres vivos. La molécula de ADN consiste en dos cadenas que se enrollan entre ellas para formar una estructura de doble hélice. Cada cadena tiene una parte central formada por azúcares (desoxirribosa) y grupos fosfato. Enganchado a cada azúcar hay una de las siguientes 4 bases: adenina (A), citosina (C), guanina (G), y timina (T). Las dos cadenas se mantienen unidas por enlaces entre las bases; la adenina se enlaza con la timina, y la citosina con la guanina. La secuencia de estas bases a lo largo de la cadena es lo que codifica las instrucciones para formar proteínas y moléculas de ARN.

ADN mitocondrial

El ADN mitocondrial es el pequeño cromosoma circular que se encuentra en la mitocondria. Las mitocondrias son orgánulos celulares donde se produce energía. Las mitocondrias, y por tanto el ADN mitocondrial, solo se heredan de la madre.

ADN no codificante

Las secuencias no codificantes de ADN no codifican para aminoácidos. La mayor parte del ADN no codificante se encuentra entre los genes en el cromosoma y no tiene función conocida. Otras secuencias de ADN no codificantes, llamadas intrones, se encuentran dentro de los genes. Parte del ADN no codificante desempeña un papel en la regulación de la expresión génica.

ADN recombinante

El ADN recombinante (rADN) es una tecnología que utiliza enzimas para cortar y unir secuencias de ADN de interés. Las secuencias de ADN recombinado se pueden colocar en unos vehículos llamados vectores que transportan el ADN hacia el lugar adecuado de la célula huésped donde puede ser copiado o expresado.

Alelo

Un alelo es cada una de las dos o más versiones de un gen. Un individuo hereda dos alelos para cada gen, uno del padre y el otro de la madre. Los alelos se encuentran en la misma posición dentro de los cromosomas homólogos. Si los dos alelos son idénticos,

[46] Las definiciones presentadas en este apartado fueron consultadas en el *Glosario de Términos Genéticos* del *Instituto Nacional de Investigación del Genoma Humano* (NHGRI). *https://www.genome.gov/glossarys/*

el individuo es homocigoto para este gen. En cambio, si los alelos son diferentes, el individuo es heterocigoto para este gen. Aunque el término alelo fue usado originariamente para describir variaciones entre los genes, ahora también se refiere a las variaciones en secuencias de ADN no codificante (es decir, que no se expresan).

ARN (ácido ribonucleico)

El ácido ribonucleico (ARN) es una molécula similar a la de ADN. A diferencia del ADN, el ARN es de cadena sencilla. Una hebra de ARN tiene un eje constituido por un azúcar (ribosa) y grupos de fosfato de forma alterna. Unidos a cada azúcar se encuentra una de las cuatro bases adenina (A), uracilo (U), citosina (C) o guanina (G). Hay diferentes tipos de ARN en la célula: ARN mensajero (ARNm), ARN ribosomal (ARNr) y ARN de transferencia (ARNt). Más recientemente, se han encontrado algunos ARN de pequeño tamaño que están involucrados en la regulación de la expresión génica.

ARN de transferencia (ARNt)

El ARN de transferencia (ARNt) es una pequeña molécula de ARN que participa en la síntesis de proteínas. Cada molécula de ARNt tiene dos áreas importantes: una región de trinucleótidos denominada anticodón y una región donde se une un aminoácido específico. Durante la traducción, cada vez que un aminoácido se añade a la cadena en crecimiento, se forma una molécula de ARNt cuyos pares de bases tienen una secuencia complementaria con la molécula del ARN mensajero (ARNm), asegurando que el aminoácido adecuado sea insertado en la proteína.

ARN mensajero (ARNm)

El ARN mensajero (ARNm) es una molécula de ARN de cadena simple, complementaria a una de las cadenas de ADN de un gen. El ARNm es una versión del ARN del gen que sale del núcleo celular y se mueve al citoplasma donde se fabrican las proteínas. Durante la síntesis de proteínas, un orgánulo llamado ribosoma se mueve a lo largo del ARNm, lee su secuencia de bases, y utiliza el código genético de traducir cada triplete de tres bases o codón, en su aminoácido correspondiente.

B

Bioinformática

La Bioinformática es una subdisciplina de la biología y las ciencias computacionales que se encarga de adquirir, almacenar, analizar y diseminar la información biológica, en gran parte correspondiente a las secuencias de ADN y aminoácidos.

C

Cáncer

El cáncer es un grupo de enfermedades caracterizadas por un crecimiento celular descontrolado. El cáncer empieza cuando una única célula muta, y se alteran los controles de regulación que mantienen a la división celular en correcto funcionamiento. Estas mutaciones pueden ser heredadas, causadas por errores en la replicación del ADN, o el resultado de la exposición a sustancias químicas nocivas. Un tumor canceroso puede propagarse a otras partes del cuerpo y, si no se trata, puede ser fatal.

Célula

Las células son los bloques estructurales básicos de los seres vivos. Todas las células se pueden clasificar en dos grupos: eucariotas y procariotas. Las eucariotas tienen núcleo y orgánulos envueltos por una membrana, mientras que las procariotas no. Las plantas y los animales están constituidas por un gran número de células eucariotas, mientras que muchos de los microbios, como las bacterias, son células individuales. Se estima que el cuerpo adulto de un humano contiene entre 10 y 100 billones de células.

Centrómero

El centrómero es la región estrecha de un cromosoma que lo separa en un brazo corto (p) y un brazo largo (q). Durante la división celular, los cromosomas se replican primero de manera que cada célula hija recibe un conjunto completo de cromosomas. A raíz de la replicación del ADN, el cromosoma queda formado por dos estructuras idénticas llamadas cromátidas hermanas, que están unidas por el centrómero.

Código genético

El código genético son las instrucciones que le dicen a la célula cómo hacer una proteína específica. A, T, C y G, son las "letras" del código del ADN; representan los compuestos químicos adenina (A), timina (T), citosina (C) y guanina (G), respectivamente, que constituyen las bases de nucleótidos del ADN. El código para cada gen combina los cuatro compuestos químicos de diferentes maneras para formar "palabras" de tres letras las cuales especifican qué aminoácidos se necesitan en cada paso de la síntesis de una proteína.

Cromátida

Una cromátida es cada una de las dos mitades idénticas de un cromosoma duplicado. Durante la división celular, en primer lugar, se duplica el cromosoma para que cada una de las células hijas reciba una dotación cromosómica completa. Después de la duplicación del ADN, el cromosoma pasa a estar compuesto por dos estructuras idénticas, llamadas cromátidas hermanas, que se unen por la zona del centrómero.

Cromosoma

Un cromosoma es un paquete ordenado de ADN que se encuentra en el núcleo de la célula. Los diferentes organismos tienen diferentes números de cromosomas. Los humanos tenemos 23 pares de cromosomas - 22 pares autosómicos, y un par de cromosomas sexuales, X e Y. Cada progenitor contribuye con un cromosoma de su par de autosomas y uno del par sexual, de manera que la descendencia obtenga la mitad de sus cromosomas de su madre y la mitad de su padre.

Cromosoma artificial bacteriano (BAC)

Un cromosoma artificial bacteriano (BAC) es una molécula de ADN utilizada para clonar secuencias de ADN en las células bacterianas (por ejemplo, E. coli). Los BAC se suelen utilizar en la secuenciación del ADN. Los segmentos de ADN de un organismo, que van de 100.000 a cerca de 300.000 pares de bases, se pueden insertar en BACs. Los BACs, con su ADN insertado, son entonces introducidos en células bacterianas. A medida que las células bacterianas crecen y se dividen, amplifican también el ADN de los BACs, que después pueden ser aislados y utilizados en la secuenciación del ADN.

E

Enzima

Una enzima es un catalizador biológico. Es una proteína que acelera la velocidad de una reacción química específica en la célula. La enzima no se destruye durante la reacción y se utiliza una y otra vez. Una célula contiene miles de diferentes tipos de moléculas de enzimas específicos para cada reacción química particular.

Epidemiología genética

La epidemiología genética es una disciplina de la Medicina relativamente nueva que trata de comprender cómo los factores genéticos interactúan con el medio ambiente en el contexto de la enfermedad en las poblaciones.

Epigenoma

Epigenoma es un término que se deriva de la palabra griega epi, que significa literalmente "*por encima*" del genoma. El epigenoma se compone de compuestos químicos que modifican, o marcan, el genoma de manera que le dice qué hacer, dónde hacerlo y cuándo hacerlo. Células diferentes tienen diferentes marcas epigenéticas. Estas marcas epigenéticas, que no forman parte del propio ADN, pueden ser transmitidas de una célula a otra durante la división celular, y de una generación a la otra.

Evolución

La evolución es el proceso mediante el cual los organismos cambian con el tiempo. Las mutaciones producen variación genética en las poblaciones y el medio ambiente interactúa con dichas variaciones seleccionando a aquellos individuos que mejor se adapten a su entorno. Los individuos mejor adaptados tienen mayor descendencia que los individuos peor adaptados. A través de un periodo largo, una especie puede evolucionar en muchas otras.

Exoma

El exoma es la parte del genoma (conjunto de moléculas de DNA) formado por los exones, los fragmentos de DNA que se transcriben para dar lugar a las proteínas. El estudio del exoma es una de las formas más completas y complejas de estudiar nuestro DNA.

Exón

Un exón es la porción de gen que codifica aminoácidos. En las células de plantas y animales, la mayoría de las secuencias de genes son alternadas por una o más secuencias de ADN llamadas intrones. Las partes de la secuencia de genes que contienen la información para producir las proteínas se llaman exones, ya que se expresan, mientras que las partes de la secuencia del

gen que no codifican se llaman intrones, porque están en medio o interfieren con los exones.

F

Farmacogenómica La Farmacogenómica es una rama de la farmacología, que utiliza el ADN y datos de la secuencia de aminacidos para aplicarlos al desarrollo de drogas y nuevas pruebas clinicas. Una aplicacion importante de la Farmacogenómica es correlacionar las variaciones geneticas individuales con la respuesta a drogas.

Fenotipo El fenotipo constituye los rasgos observables de un individuo, tales como la altura, el color de ojos, y el grupo sanguíneo. La contribución genética al fenotipo se llama genotipo. Algunos rasgos son determinados en gran medida por el genotipo, mientras que otros rasgos están determinados en gran medida por factores ambientales.

G

Gen El gen es la unidad física básica de la herencia. Los genes se transmiten de los padres a la descendencia y contienen la información necesaria para precisar sus rasgos. Los genes están dispuestos, uno tras otro, en estructuras llamadas cromosomas. Un cromosoma contiene una única molécula larga de ADN, sólo una parte de la cual corresponde a un gen individual. Los seres humanos tienen aproximadamente 20.000 genes organizados en sus cromosomas.

Genoma El genoma es el conjunto de instrucciones genéticas que se encuentra en una célula. En los seres humanos, el genoma consiste en 23 pares de cromosomas, que se encuentran en el núcleo, así como un pequeño cromosoma que se encuentra en las mitocondrias de las células. Cada conjunto de 23 cromosomas contiene aproximadamente 3,1 mil millones de bases de la secuencia de ADN.

Genómica La genómica se refiere al estudio del genoma completo de un organismo, mientras que la genética se refiere al estudio de un gen en concreto.

Genómica de poblaciones

Genómica de poblaciones es la aplicación de las tecnologías genómicas para entender las poblaciones de organismos. En los seres humanos, la genómica de la población general se refiere a la aplicación de tecnología en la búsqueda para entender cómo los genes contribuyen a nuestra salud y bienestar.

Genotipo

Un genotipo es la colección de genes de un individuo. El término también puede referirse a los dos alelos heredados de un gen en particular. El genotipo se expresa cuando la información codificada en el ADN de los genes se utiliza para fabricar proteínas y moléculas de ARN. La expresión del genotipo contribuye a los rasgos observables del individuo, lo que se denomina el fenotipo.

H

Haplogrupo

Un haplogrupo es, en el estudio de la evolución molecular, un grupo grande de haplotipos, que son series de alelos en lugares específicos de un cromosoma.

Haplotipo

Un haplotipo es un conjunto de variaciones del ADN, o polimorfismos, que tienden a ser heredados juntos. Haplotipo se puede referir a una combinación de alelos o a un conjunto de polimorfismos de nucleótido sencillo (SNPs) que se encuentran en el mismo cromosoma. Información acerca de distintos haplotipos está siendo recopilada por el Proyecto Internacional HapMap y utilizada para investigar la influencia de los genes en enfermedades.

Histona

Una histona es una proteína que proporciona soporte estructural a un cromosoma. Para que las larguísimas moléculas de ADN quepan en el núcleo celular, se envuelven alrededor de complejos de histonas, dando al cromosoma una forma más compacta. Algunas variantes de las histonas están asociadas con la regulación de la expresión génica.

HUGO

El Proyecto HUGO (genoma humano) es un proyecto el cual tiene como principal objetivo el conocimiento de la secuencia de bases y pares químicos que el ADN e identificar y cartografiar todos los genes del genoma humano desde un punto de vista físico y funcional.

I

Ingeniería genética

La ingeniería genética es el proceso de la utilización de la tecnología del ADN recombinante (ADNr) para alterar la composición genética de un organismo. Tradicionalmente, los seres humanos han manipulado indirectamente los genomas mediante el control de la reproducción, así como seleccionando aquella descendencia que tenga las características deseadas. La ingeniería genética implica la manipulación directa de uno o más genes. Lo más común es que un gen de otra especie se introduzca en el genoma de un organismo para producir el fenotipo deseado.

Iniciador o cebador

Un iniciador o cebador es una secuencia corta de ADN de cadena simple que se utiliza en una reacción en cadena de la polimerasa (PCR). En el método PCR se emplea un par de cebadores para hibridar con el ADN de la muestra y definir la región del ADN que será amplificada. También se les conoce como oligonucleótidos.

Intrón

Un intrón es una parte del gen que no codifica ningún aminoácido. En las células vegetales y animales, la mayoría de las secuencias que codifican para los genes están partidas por uno o más intrones. Las zonas de la secuencia del gen que se expresan en las proteínas se llaman exones porque se expresan, mientras que aquellas que no lo hacen se denominan intrones por encontrarse entre los exones.

L

LD

En genética se denomina desequilibrio del ligamiento a la propiedad de algunos genes de las poblaciones genéticas de no segregar de forma independiente, esto es, poseen una frecuencia de recombinación menor del 50%. Esto suele deberse a que los dos loci implicados se encuentran en el mismo cromosoma, lo que imposibilita su transferencia a la progenie de manera aleatoria con la separación de los cromosomas en anafase.

Ligamiento

El ligamiento es la asociación de genes u otras secuencias cercanas del ADN en el mismo cromosoma. Cuanto más cerca están dos genes en el cromosoma, mayor es la posibilidad de que se hereden juntos.

Locus

Un locus es el lugar específico del cromosoma donde está localizado un gen u otra secuencia de ADN, como su dirección genética. El plural de locus es "loci".".

M

Marcador genético

Un marcador genético es un segmento de ADN con una ubicación física conocida en un cromosoma. Los marcadores genéticos pueden ayudar a vincular una enfermedad hereditaria con el gen responsable. Los segmentos de ADN que se encuentran cerca en un cromosoma tienden a heredarse juntos.

Marco abierto de lectura

Marco abierto de lectura es una porción de una molécula de ADN que cuando se traduce a los aminoácidos, no contiene codones de terminación. El código genético lee secuencias de ADN en grupos de tres pares de bases, esto significa que, en una molécula de ADN de doble hebra, hay 6 posibles sentidos en los que pueden abrirse marcos de lectura --tres en dirección hacia adelante y tres en reverso. Un marco abierto de lectura larga es probable que sea parte de un gen.

Medicina Personalizada

La medicina personalizada es una práctica emergente de la medicina que utiliza el perfil genético de un individuo para guiar las decisiones tomadas en relación con la prevención, diagnóstico y tratamiento de la enfermedad. El conocimiento del perfil genético de un paciente puede ayudar a los médicos seleccionar la medicina o la terapia adecuada, así como administrar la dosis o el régimen adecuados. La medicina personalizada está avanzando gracias a los datos del Proyecto Genoma Humano.

N

Nucleosoma

El nucleosoma es la unidad básica de repetición de la cromatina eucariótica. En una célula humana, cerca de dos metros de ADN deben ser empaquetados en un núcleo con un diámetro inferior a un cabello humano. Un nucleosoma se compone de alrededor de 150 pares de bases de ADN enrolladas alrededor de un núcleo de histonas. Los nucleosomas se organizan como cuentas de un collar las cuales, a su vez, son plegadas sobre sí mismas repetidas veces para formar un cromosoma.

Nucleótido

Un nucleótido es la pieza básica de los ácidos nucleicos. El ARN y el ADN son polímeros formados por largas cadenas de nucleótidos. Un nucleótido está formado por una molécula de azúcar (ribosa en el ARN o desoxirribosa en el ADN) unido a un grupo fosfato y una base nitrogenada. Las bases utilizadas en el ADN son la adenina (A), citosina (C), guanina (G) y timina (T). En el ARN, la base uracilo (U) ocupa el lugar de la timina.

P

Pares de base

Un par de bases es un par de bases químicas que interaccionan entre ellas. Podemos imaginar que la doble hélice de ADN es como una escalera de mano, donde los pasamanos son las dos hebras enrolladas entre sí. La unión entre los pares de bases corresponde al peldaño de la escalera. Cada hebra está formada por la alternancia de un azúcar (desoxirribosa) y un grupo fosfato. En cada azúcar, hay anclada una de las cuatro bases nitrogenadas: adenina (A), citosina (C), guanina (G) o timina (T). Las dos hebras se mantienen juntas gracias a los puentes de hidrógeno entre las bases complementarias, es decir, la adenina con la timina, y la citosina con la guanina.

Polimorfismo

Un polimorfismo implica una de dos o más variantes de una secuencia particular de ADN. El tipo más común de polimorfismo implica la variación en un solo par de bases. Los polimorfismos también pueden ser de mucho mayor tamaño implicando largos tramos de ADN. Los llamados polimorfismos de nucleótido sencillo, o SNP (por sus siglas en inglés y pronunciado "esnip"), están siendo estudiados por los científicos para ver su correlación en el genoma humano con enfermedades, respuesta a los fármacos, y otros fenotipos.

Promotor

El promotor es una secuencia de ADN necesaria para convertir un gen en activado o desactivado. El proceso de transcripción se inicia en el promotor. Generalmente se encuentran cerca del comienzo de un gen, el promotor tiene un sitio de unión para la enzima que se utiliza para hacer una molécula ARN mensajero (ARNm).

Proteína

Las proteínas son una clase importante de moléculas que se encuentran en todas las células vivas. Una proteína se compone de una o más cadenas largas de aminoácidos, cuya secuencia corresponde a la secuencia de ADN del gen que la codifica. Las proteínas desempeñan gran variedad de funciones en la célula, incluidas estructurales (citoesqueleto), mecánicas (músculo), bioquímicas (enzimas), y de señalización celular (hormonas). Las proteínas son también parte esencial de la dieta.

R

Reacción en cadena de la polimerasa (PCR)

La reacción en cadena de la polimerasa (PCR) es una técnica de laboratorio utilizada para amplificar secuencias de ADN. El método utiliza secuencias cortas de ADN llamados cebadores para seleccionar la parte del genoma a amplificar.

Repetición en tándem

Una repetición en tándem es una secuencia de dos o más pares de bases de ADN que se repite de tal manera que las repeticiones se encuentran uno al lado del otro en el cromosoma. Repeticiones en tándem están generalmente asociadas con el ADN no codificante. En algunos casos, el número de veces que se repite la secuencia de ADN es variable. Dicha variabilidad de repeticiones en tándem se puede utilizar como una "huella" genética.

Replicación de ADN

La replicación del ADN es el proceso mediante el cual se duplica una molécula de ADN. Cuando una célula se divide, en primer lugar, debe duplicar su genoma para que cada célula hija contenga un juego completo de cromosomas.

S

SNP

Los polimorfismos de nucleótido único (SNP) son un tipo de polimorfismo que producen una variación en un solo par de bases. Los científicos están estudiando cómo los polimorfismos de nucleótido único, o SNPs (pronunciado "snips"), en el genoma humano se correlacionan con enfermedades, con la respuesta de los fármacos, y con otros fenotipos.

T

Tecnología de microarrays (chips de ADN o ARN)

La tecnología de microarrays es una tecnología en desarrollo para estudiar la expresión de muchos genes a la vez. Consiste en colocar miles de secuencias génicas en lugares determinados sobre un portaobjetos de vidrio llamado chip. Una muestra que contiene ADN o ARN se pone en contacto con el chip. El apareamiento de las bases complementarias entre la muestra y las secuencias de genes en el chip produce una cantidad de luz que se puede medir. Las áreas del chip que producen luz identifican los genes que se expresan en esa muestra.

Telómero

Un telómero es el final de un cromosoma. Los telómeros son secuencias repetitivas de ADN no codificante del cromosoma que protegen de cualquier daño. Cada vez que una célula se divide, los telómeros se acortan. Con el tiempo, los telómeros se vuelven tan cortos que la célula ya no puede dividirse.

Traducción

La traducción es el proceso de traducir la secuencia de una molécula de ARN mensajero (ARNm) a una secuencia de aminoácidos durante síntesis de proteínas. El código genético se describe la relación entre la secuencia de pares de bases en un gen y la secuencia correspondiente de aminoácidos que codifica.

Transcripción

Transcripción es el proceso por el cual se genera una copia de RNA a partir la secuencia de un gene. Esta copia, llamada una molécula de ARN mensajero (ARNm), deja el núcleo de la célula y entra en el citoplasma, donde dirige la síntesis de la proteína, que codifica.

Transgénico

Transgénico significa que una o más secuencias de ADN de otra especie han sido introducidos por medios artificiales. Los animales transgénicos por lo general se producen a partir de una pequeña secuencia de ADN extraño que se inyecta en un óvulo fecundado o embrión en desarrollo. Las plantas transgénicas se pueden hacer mediante la introducción de ADN extraño en una variedad de diferentes tejidos.

V

Variabilidad genética La variabilidad genética se refiere a la diversidad en las frecuencias de los genes. La variabilidad genética puede referirse a las diferencias entre individuos o las diferencias entre poblaciones. Las mutaciones son la causa fundamental de la variabilidad genética, pero mecanismos tales como la reproducción sexual y la deriva genética también contribuyen a la misma.

Variación Cambios en la secuencia respecto a la secuencia de referencia.

Variación en el número de copias (VCN) La variación en el número de copias es cuando un número de copias de un gen particular cambia de un individuo a otro. Tras la finalización del Proyecto Genoma Humano se hizo evidente que el genoma experimenta ganancias y pérdidas de material. La medida en que la variación del número de copias contribuye a la enfermedad humana no se conoce todavía. Desde hace tiempo se ha visto que algunos cánceres están asociados con elevados números de copias de genes específicos.