Máster Universitario en Ingeniería y Tecnología de Sistemas Software

Departamento de Sistemas Informáticos y Computación

Exploración de Bases de Datos Genómicas Dirigida por Modelos Conceptuales

Septiembre 2018

Vanessa Alexandra Solis Cabrera
vasocab@posgrado.upv.es

Director: Óscar Pastor López
opastor@pros.upv.es

Director experimental: Ana León
aleon@pros.upv.es

RESUMEN

Cada día los doctores, genetistas e investigadores enfrentan grandes retos al momento de buscar información y querer sintetizar los resultados obtenidos luego de una consulta de información médica en la web. No obstante, internamente quedan las dudas sobre: <<Se obtuvieron los resultados esperados>> o <<Existe información adicional en otros lugares que no se consideraron>> o <<La información que no se consideró es relevante>> entre otras. Pues muchas de las veces lo que se espera es abarcar todas las fuentes de información para generar así un reporte médico completo. De manera similar sucede en el ámbito del genoma humano, donde se pueden encontrar cientos de páginas con bases de datos genómicas. Varios de estos sitios son desconocidos, ya sea porque son nuevos o no han sido publicitados.

Ahora bien, la interrogante del investigador en el momento de sintetizar información es <<¿Cómo unifico los resultados, si en las consultas realizadas he obtenido diferente información?>>, <<¿Cuáles de los datos son iguales o tienen igual significado?>>. La aplicación del modelado conceptual aplicado en este caso al ámbito médico permite considerar los diferentes procesos y detallar las tareas y actividades que se desarrollan. Pues ahí es donde el modelado conceptual juega un rol muy importante, puesto que, luego de abstraer los elementos que se desea obtener se permite encontrar coincidencias desde varias fuentes y que estas apunten a un mismo concepto.

Es así que, el modelado del genoma humano es una pieza fundamental que permite considerar las entidades involucradas y sus relaciones. Por este motivo, el presente trabajo de fin de máster pretende ser una herramienta de ayuda para las personas involucradas con el área del genoma humano. Al crearse un mapeo, es decir al establecerse los vínculos entre la información las bases de datos genómicas contrastadas con cada uno de los elementos del Esquema Conceptual del Genoma Humano (ECGH), se establecen los enlaces necesarios para que los usuarios tengan un soporte sus búsquedas. Este trabajo presenta el desarrollo de la exploración de las bases de datos genómicas que se han encontrado en listados avalados por Institutos de investigación en el área genómica. Se indica el proceso de verificación de los sitios que se han encontrado, puesto que algunos sitios han sufrido cambios en los servidores o simplemente ya dejaron de funcionar. Así también se exponen las tareas de depuración que se ha generado, debido a que cada una de las bases de datos genómicas presentan diferentes estructuras, organizaciones de la información, incluso en algunos de los casos se utiliza nomenclatura distinta a la que habitualmente el investigador está acostumbrado a encontrar. Posteriormente se presenta el mapeo de cada base de datos genómica con los elementos del ECGH. Finalmente se muestra los resultados obtenidos con estadísticas que se pudieron establecer en la exploración de las bases de datos genómicas y un interfaz resultado como herramienta de ayuda para el usuario.

Para finalizar, se exponen los problemas encontrados en el desarrollo de este trabajo, así como también las conclusiones a las que se han podido llegar. Se indican los posibles trabajos futuros que se pueden realizar a partir de este trabajo.

Palabras claves: Bases de datos genómicas, Modelo conceptual, Esquema conceptual del genoma humano.

ABSTRACT

Doctors, geneticists and researchers face huge challenges every day; especially at the time of looking for information and to synthesize the results obtained after checking medical information online. However, there are certain doubts regarding this issue; for example, <<The expected results were obtained>> or <<There is additional information in different places that was not taken into account>> or <<The information that was not included was relevant>>, among others. Most of the time, it is expected to encompass all the sources to generate a complete medical report. It happens in a similar way in the human genome field where hundreds of pages with data bases are found. Many of those are unknown since they are either new or have been advertised.

So the questions for the researcher at the time of synthesizing information is, How to unify results if the searches done have brought different information? Which pieces of information are the same or have the same meaning? In this case, the application of the conceptual model in the medical field allows to consider different processes to detail the tasks and activities that are developed. It is there where the conceptual model plays an important role, given that once the elements are abstracted and that the desired information is obtained, it is possible to find information from many sources so that they match one single concept.

Thereof is the fact that the human genome modeling is a fundamental piece that allows to consider the different entities and their relationships. For this reason, this master's degree work aims to be a tool for those professionals involved in the human genome field. Setting up a map, that is to say, establishing links among the genomic data bases contrasted to each one of the elements from the Human Genomic Conceptual Scheme (HGCS), the necessary links are established so that the users find support when searching information. This work presents the development of the genomic data bases exploration that has been found on validated sources; that is, by research institutions in the area of genetics. The verification process of the sites that were found is shown because several sites have been changed in the servers or they simply stopped working. In addition, the debugging tasks are mentioned in this work because each one of the genomic data bases presents different structures, information organization, and in some cases, they used different naming compared to the one the researcher is used to find. After that, a genomic data base mapping is presented with the HGCS. Finally, the obtained results are shown in this work, including statistics that were established through the genomic data bases exploration and the interface that was the result as a tool for the user.

In a nutshell, the problems found during the development of this work are shown along with the conclusions that arose. Possible future researches on the bases of this work are indicated.

Keywords: Genomic database, Conceptual Model, Human Genomic Conceptual Scheme.

AGRADECIMIENTOS

Quiero agradecer primeramente a Dios quien me ha guiado en este caminar y así poder culminar una nueva etapa profesional llena de grandes y gratos momentos.

A mi esposo por ser la persona que con muchos cuidados y paciencia ha tenido gestos muy lindos en este año. Aprendimos que caminar juntos es una tarea fácil, pero que empieza a complicarse cuando en ella se incluyen los factores deberes, tareas y sobre todo exámenes. Gracias por todo Juanjo!!!

A mis Padres Marcelo y Rosa quienes cada día estuvieron presentes desde el amanecer hasta el anochecer por cada detallito que nos pudiese faltar. Ese tiempo realmente es oro para mí ya que fortaleció mi espíritu y mis vínculos con ustedes. Cada caída tenía una frase de motivación, cada logro frases de éxito y cada éxito frases de que aún falta dar lo mejor de mí. Gracias por todo lo que hacen, a pesar de la distancia, este no ha sido un impedimento para estar siempre conectados así sea de manera virtual.

Zaidita mil gracias por todos esos momentos de alegrías y sustos que pudimos vivir en este largo, pero a su vez corto tiempo en Valencia con usted. Todos esos momentos quedarán impregnados para contar a las generaciones venideras.

Gracias a toda mi familia en Cuenca Ecuador que a pesar de que no los nombre uno a uno siempre están en mis pensamientos por todas esas palabras de ánimo y aliento. Por estar pendientes de cómo es acá el ambiente, el clima, la comida, la playa en fin cada detalle que experimentaba. Si bien el no poder compartir todo en persona con ustedes es difícil, pues me enseñó a que mediante historias o mensajitos compartiendo mis vivencias también llenaría sus mentes y corazones.

Óscar gracias por encaminarme nuevamente al apasionante mundo de la medicina. Gracias por darme la oportunidad de compartir conocimientos en pro de mejoras no solo personales sino en bien de la sociedad. Gracias por abrirme las puertas de un grupo de investigación como lo es el PROS, en donde comprender el funcionamiento de la vida desde ámbito tecnológico marca muchos retos por cumplir.

Ana L. cada uno de los momentos que dedicaste en las revisiones y enseñanzas sobre el ámbito genómico fue de gran ayuda en mi crecer profesional. Gracias por compartir tu conocimiento eres una gran persona! Así mismo José R. tus consejos y explicaciones fueron muy oportunas para poder completar mi TFM.

Gracias Simran, Andriy y Joan excelentes compañeros que pude conocer en este máster. Todos esos instantes cuando los conocimientos no solo quedaban en cada uno de nuestros apuntes sino pudimos compartirlos creciendo como profesionales y amigos cheveres.

Luli V. gracias por el tiempo dedicado a explicarme detallitos de genética que parecían en otro idioma pero al final fue juego de palabras. María Isabel gracias, ahora si prometo afinar mi inglés!

No por expresarlo al final son menos importantes, pero sin la compañía de ustedes la estancia pudo ser dura y complicada, gracias por ser la familia en Valencia. Kary, Angélica, Miguel, Orlando, Stalin, Joaquín, Joel, Nicole, Eugenio, Lore, Leo, toda su amistad ha caído en campo fértil y espero perdure muchos años más.

ÍNDICE DE CONTENIDO

RESUMEN 2
ABSTRACT 3
AGRADECIMIENTOS 4
ÍNDICE DE CONTENIDO 5
ÍNDICE DE TABLAS 6
ÍNDICE DE ILUSTRACIONES 6
GLOSARIO DE TÉRMINOS 8
1 INTRODUCCIÓN 10
1.1 MOTIVACION 10
1.2 PLANTEAMIENTO DEL PROBLEMA 11
1.3 OBJETIVOS 11
1.4 METODOLOGÍA 12
1.5 ESTRUCTURA DEL TRABAJO 14
2 MODELO CONCEPTUAL 15
2.1 COMPONENTES BÁSICOS DEL MODELO CONCEPTUAL GENÓMICO 16
3 BASES DE DATOS GENÓMICAS 19
3.1 BASES DE DATOS GENÓMICAS 19
3.1.1 OXFORD ACADEMIC 21
3.1.2 NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION 22
3.1.3 HUMAN GENOME VARIATION SOCIETY 23
3.1.4 HEALTH SCIENCES LIBRARY SYSTEM 25
4 MAPA DE BASE DE DATOS GENÓMICAS SOBRE EL ESQUEMA CONCEPTUAL DEL GENOMA HUMANO 27
4.1 CRITERIOS DE SELECCIÓN 27
4.2 SELECCIÓN DE BASES DE DATOS GENÓMICAS 28
4.2.1 EXTRACCIÓN DE LISTADOS DE BASES DE DATOS 28
4.2.2 TECNOLOGÍAS EMPLEADAS 32
4.2.3 LIMPIEZA DE BASES DE DATOS 32
4.3 TRAZABILIDAD ENTRE MODELO CONCEPTUAL Y BD 34
4.4 PRESENTACIÓN DE RESULTADOS 43
4.4.1 TECNOLOGÍAS EMPLEADAS 43

4.4.2 SITIO WEB 44
4.4.3 DATOS ESTADÍSTICOS 48
5 CONCLUSIONES 56
5.1 PROBLEMAS ENCONTRADOS 56
5.2 CONCLUSIONES 63
5.3 TRABAJOS FUTUROS 64
REFERENCIAS BIBLIOGRÁFICAS 65
ANEXOS 69
ANEXO 1 – ESQUEMA CONCEPTUAL DEL GENOMA HUMANO 69
VISTA ESTRUCTURAL 69
VISTA DE TRANSCRIPCIÓN 70
VISTA DE VARIACIONES 72
VISTA DE RUTAS METABÓLICAS O PATHWAYS 75
VISTA DE FUENTES DE DATOS Y BIBLIOGRAFÍA 77

ÍNDICE DE TABLAS

TABLA 1 CATEGORÍAS DE DB EN NAR 21
TABLA 2 CATEGORÍAS DE DB EN NCBI 23
TABLA 3 CATEGORÍA DE DB DE HGVS 24
TABLA 4 CATEGORÍAS DE BD DE HSLS 25
TABLA 5 PRESELECCIÓN DE BASES DE DATOS GENÓMICAS 33
TABLA 6 CÓDIGOS DE RESPUESTA EN COMPROBACIÓN DE SITIOS WEB 33
TABLA 7 MAPEO DE ELEMENTOS DEL ECGH Y BD GENÓMICA 38
TABLA 8 MAPEO DE ELEMENTOS DEL ECGH Y BD GENÓMICA 41
TABLA 9 OTRAS ESPECIES 49
TABLA 10 ERRORES EN CARGA DE SITIOS WEB DE DB GENÓMICAS 51
TABLA 11 ERRORES ENCONTRADOS EN SITIOS CON CONEXIÓN 53
TABLA 12 DB GENÓMICAS QUE PERMITEN DESCARGA DE RECURSOS (ESPECIE HUMANA) 54
TABLA 13 DB GENÓMICAS QUE INCORPORAN APIS (ESPECIE HUMANA) 54
TABLA 14 OTRAS BASES DE DATOS GENÓMICAS 55

ÍNDICE DE ILUSTRACIONES

ILUSTRACIÓN 1 METODOLOGÍA DE INVESTIGACIÓN UTILIZADA 13
ILUSTRACIÓN 2 DE LOS MODELOS CONCEPTUALES AL CÓDIGO: UNA PERSPECTIVA SE Y UNA PERSPECTIVA DE COMPRENSIÓN DE LA VIDA 15
ILUSTRACIÓN 3 VISTAS DEL MODELO CONCEPTUAL HOLÍSTICO 16
ILUSTRACIÓN 4 MODELO CONCEPTUAL A CONSIDERARSE 18
ILUSTRACIÓN 5 ORIGEN DE BASES DE DATOS GENÓMICAS 20
ILUSTRACIÓN 6 EXTRACCIÓN DE DATOS DE OXFORD JOURNAL 29

Ilustración 7 Revisión de HGVS 30
Ilustración 8 Extracción de datos de HGVS 30
Ilustración 9 Extracción de datos de HSLS 31
Ilustración 10 Selección de cada una de las categorías de HSLS 31
Ilustración 11 Extracción de datos de NCBI 32
Ilustración 12 Flujo para el mapeo de las BDs y ECGH 34
Ilustración 13 Base de datos no disponible 35
Ilustración 14 Base de datos sin acceso 35
Ilustración 15 Extracción de datos informativos 36
Ilustración 16 Identificación de elementos para mapeo 36
Ilustración 17 Mapeo de información de BD y ECGH 37
Ilustración 18 Archivo CSV de registro del mapeo 38
Ilustración 19 Extracción básica de información 39
Ilustración 20 Búsqueda de atributos en diccionario de datos 39
Ilustración 21 Búsqueda de información por gen 40
Ilustración 22 Búsqueda por nombre de enfermedad 40
Ilustración 23 Mapeo de información con el ECGH 42
Ilustración 24 Trazabilidad mediante elementos de página web 43
Ilustración 25 Diagrama de base de datos 44
Ilustración 26 Herramienta para la visualización de resultados 45
Ilustración 27 Visualizador creado para búsquedas 45
Ilustración 28 Visualizar BD genómicas por Vistas 46
Ilustración 29 Visualizar BD genómicas por Clases 46
Ilustración 30 Visualizar BD genómicas por Atributos 47
Ilustración 31 Visualizar contenido de BD genómica 48
Ilustración 32 BD genómicas por especie 49
Ilustración 33 Acceso a BD genómicas 51
Ilustración 34 Causas de no conexión 52
Ilustración 35 Acceso a las BD genómicas: Especie Humana 52
Ilustración 36 Causas de no acceso a las BD genómicas: Especie Humana 53
Ilustración 37 BD agrupados por vistas 54
Ilustración 38 Ejemplo problema 1 57
Ilustración 39 Ejemplo problema 2 57
Ilustración 40 Ejemplo problema 3 58
Ilustración 41 Ejemplo problema 4 58
Ilustración 42 Ejemplo problema 5 59
Ilustración 43 Ejemplo problema 6 60
Ilustración 44 Ejemplo problema 7 61
Ilustración 45 Ejemplo problema 8 62
Ilustración 46 Ejemplo problema 9 62

GLOSARIO DE TÉRMINOS

A

ADN
: Acido Dexorribonucleico, consiste en dos moléculas parecidas a cadenas (polinucleótidos) que se tuercen alrededor de la otra para formar la clásica doble hélice., 15, 16, 20, 21, 22, 23, 24, 26, 57, 61, 69, 70, 71, 72, 73, 74, 75

API
: Application Program Interface, Interfaz de desarrollo de aplicación, 11, 46, 54

C

cromosoma
: Un cromosoma es un paquete ordenado de ADN que se encuentra en el núcleo de la célula., 24, 39, 69, 70, 72, 74, 77

E

ECGH
: Esquema conceptual del genóma humano, 2, 10, 11, 12, 13, 14, 16, 17, 20, 27, 34, 36, 37, 38, 39, 40, 41, 42, 43, 44, 47, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64

G

GenBank
: GenBank es la base de datos de secuencias genéticas NIH, una colección anotada de todas las secuencias de ADN disponibles públicamente., 22
: GenBank® es la base de datos de secuencias genéticas NIH, una colección anotada de todas las secuencias de ADN disponibles públicamente., 22

génica
: La expresión génica es el proceso mediante el cual la información codificada en un gen se utiliza para dirigir el montaje de una molécula de proteína., 25, 73

genoma
: El genoma es el conjunto de instrucciones genéticas que se encuentra en una célula. En los seres humanos, el genoma consiste de 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., 2, 10, 11, 12, 16, 19, 22, 23, 27, 28, 32, 56, 63, 64, 69, 73, 74

H

HUGO
: Organización del Genoma Humano, 23

I

IFHGS
: Federación Internacional de Sociedades de Genética Humana, 23

ISO/IEC 25010
: Organización Internacional de Normalización/Comisión Electrotécnica Internacional - International Organization for Standardization/International Electrotechnical Commission, 14

J

JSON
: JavaScript Object Notation (notación de objeto de JavaScript) es un formato de texto ligero para el intercambio de datos., 11

M

microarray
: La tecnología de microarrays es una tecnología en desarrollo para estudiar la expresión de muchos genes a la vez., 25

N

NGS
: Next Generation Sequencing, 19

nucleótido
: Un nucleótido es la pieza básica de los ácidos nucleicos. El ARN y el ADN son polímeros formados por largas cadenas de nucleótidos., 73

P

pathway
: También conocido como ruta metabólica es una serie de reacciones consecutivas catalizadas por

un enzima que produce compuestos intermedios y finalmente un producto o productos, 76

proteína
Son macromoléculas formadas por cadenas lineales de aminoácidos., 57, 70, 72

R

RNA
ácido ribonucleico (ARN) es una molécula similar a la de ADN. A diferencia del ADN, el ARN es de cadena sencilla., 21, 26, 50, 70

S

SIGe
Sistema de Información Genómico., 11, 64

T

TXT
Extensión de archivo .txt es un archivo de texto simple, texto sencillo o texto sin formato., 11

V

VCF
Variant Call Format es un fichero de texto que se usa en Bioinformática para almacenar variaciones de la secuencia de genes y su información., 11

X

XML
Extensible Markup Language (Lenguaje de Marcado Extensible) y es una especificación de recomenación W3C como lenguaje de marcado de propósito general., 11

1 INTRODUCCIÓN

1.1 MOTIVACION

El número de bases de datos biológicas disponibles para consulta pública está creciendo rápidamente tanto en lo público como en lo privado [1] [2]. Dichas bases de datos cubren distintas partes de la biología humana, desde la secuencia genética hasta la farmacoterapéutica. La unión de todo este conocimiento bajo una única perspectiva global permite la aparición de nuevos paradigmas de prevención, diagnóstico y tratamiento como por ejemplo la Medicina de Precisión [3] [4]. Es así que cada vez se requiere de una homogenización de los elementos que se tienen en las bases de datos genómicas [4].

El tratamiento de la información genómica por naturaleza requiere de un gran esfuerzo y trabajo. No obstante, cada uno de los investigadores sean genetistas o personal que trabaja en el tratamiento de la información requiere de una gran inversión de tiempo para poder encontrar, clasificar y trabajar con la información requerida, esto debido a la diversidad de fuentes de datos disponibles en el mercado ya sean de manera pública o privada, pero sobre todo por la estructura que cada una de ellas mantiene [4]. Actualmente es muy importante la disponibilidad de información que se pueda obtener de las diferentes bases de datos genómicas, ya sea con fines investigativos o vinculación con sistemas de salud o sistemas genómicos [5] [6]. No obstante, se debe tener en cuenta que cada institución ha publicado información en cada uno de los repositorios en la web bajo su propia estructura y organización de información, es decir cada autor y/o responsable de los repositorios ha seleccionado la información que ha sido valiosa o ha servido como fuente para sus investigaciones y posteriormente, por lo que se requiere tener un criterio de qué tipo de información mínima de datos se requiere a la hora de integrar diferentes fuentes.

El esquema conceptual proporciona una visión global e integradora de los distintos componentes que forman parte para una investigación del genoma humano. De esta manera se tiene la relación de los diferentes conceptos que están involucrados dentro del dominio del problema que en este caso sería el del genoma humano [7] [8]. A lo largo del tiempo han existido diferentes esquemas conceptuales desarrollados, por ejemplo el realizado por *Paton* en el 2000 [9] [10], *Ram* y *Wei* orientado a las proteínas en 3D [11], o la propuesta de Bernasconi en 2017 [12], de esta manera uno de los principales beneficios principales en cada uno de ellos ha sido la representación de los conceptos. Es así que el Esquema Conceptual del Genoma Humano (ECGH) permite expresar las relaciones entre cada uno de los elementos abstraídos del genoma, en el cual ya no solamente se base en conceptos sino en el contraste de elementos pertenecientes a los diferentes repositorios o bases de datos públicos. Como resultado de este contraste se genera un mapa de bases de datos genómicas sobre ECGH que permite la detección de la fuente de cada uno de los diferentes elementos del modelo conceptual con su origen (base de datos). Con la aplicación del ECGH se presenta una gran ventaja ya que está enfocado directamente al estudio del genoma humano, si bien el planteado por Paton fue una aproximación, esta versión no consideró diferentes elementos que con el paso del tiempo se vieron relevantes, así mismo el planteado por Ram y Wei consideró proteínas y finalmente el ahora expuesto permite tener una amplia visión para el genoma humano.

1.2 PLANTEAMIENTO DEL PROBLEMA

A medida que el conocimiento sobre la biología humana ha ido evolucionando, también lo han hecho el número de fuentes de datos que lo almacenan. Sin embargo, estos repositorios se han ido creando con propósitos muy concretos sin tener en cuenta la conexión de la información entre ellos. Esta situación ha generado problemas como inconsistencias de datos, redundancia e incluso discrepancias a la hora de representar un mismo concepto. El Esquema Conceptual del Genoma Humano (ECGH) permite unificar bajo una misma perspectiva todo el conocimiento generado en los distintos ámbitos de estudio de la genética. Así mismo, el ECGH también sirve como base para la creación de un Sistema de Información Genómico (SIGe) que permite el análisis y explotación de la información almacenada con fines diagnósticos. Pero para poder poblar este SIGe con información relevante para el diagnóstico clínico es importante resolver los problemas de integración que surgen al unir información proveniente de distintos repositorios. Es así que el primer paso será la identificación de las bases de datos genómicas que posean información y se relacionen con el ECGH.

Partiendo de una serie de bases de datos que luego de diferentes procesos de depuración y verificación, han sido seleccionadas por su calidad y su utilidad para este trabajo, se realizará un mapping entre el contenido de cada una de ellas y el ECGH. Para ello será necesario conocer:

- La representación de la información de acuerdo al ECGH, en el cual se da a conocer cada uno de los aspectos y elementos que conforman el ECGH.
- Las distintas formas de acceso a la información que ofrece cada base de datos (API, VCF, XML, TXT, JSON, entre otras).
- La correspondencia entre la información requerida por el ECGH y la información proporcionada por cada base de datos, es decir el enlace que se presenta entre cada uno de los elementos del ECGG y la información presentada en cada base de datos genómica.

La determinación de correspondencias entre las bases de datos y el ECGH, utilizando una especificación estricta, permitirá conocer en qué bases de datos genómicas se pueden crear scripts específicos para la extracción de información. Así mismo la identificación de errores comunes en la integración de información permitirá el desarrollo de estrategias globales para su resolución.

1.3 OBJETIVOS

El manejo de diferentes fuentes de información en el ámbito genómico cada día resulta más laborioso, por lo que este trabajo se enfocará como objetivo general en la exploración de las diferentes bases de datos genómicas dirigida por modelos conceptuales.

Los objetivos específicos se centrarán básicamente en:

- Seleccionar las bases de datos genómicas que correspondan al genoma humano y que estén habilitadas y funcionando.

- Especificar la correspondencia entre los atributos requeridos por el ECGH para una determinada área de conocimiento y cada una de las bases de datos seleccionadas.

- Identificar los problemas comunes a la hora de integrar información heterogénea.

- Generar una herramienta que permita visualizar las bases de datos genómicas orientadas al genoma humano en correspondencia con el ECGH.

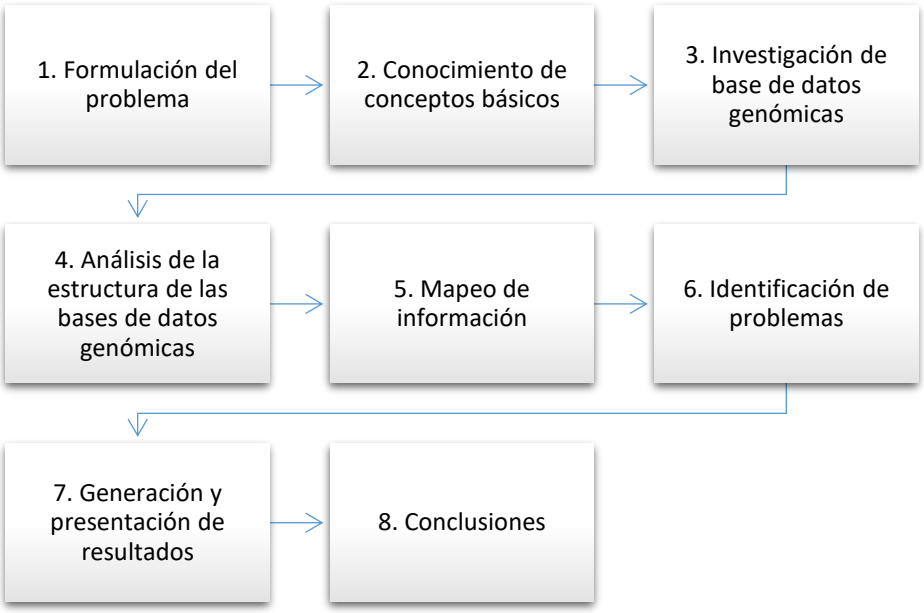
1.4 METODOLOGÍA

La metodología de trabajo a seguir para lograr la consecución de los objetivos de este trabajo consta de las siguientes etapas (Ilustración 1) [13]. Se ha seguido Design Science que ha permitido identificar cada uno de los elementos a desarrollarse en base al contexto del genoma humano, así mismo la elaboración de cada una de las actividades a realizarse.

Bajo las preguntas de investigación en base a Design Science se han generado las siguientes:

- ¿Cuál es el problema que se desea abordar? Cuya respuesta se solventa con la formulación del problema.
- ¿Qué conocimientos debo tener para la elaboración de este trabajo? La presentación del capítulo 2 y 3 reflejará la información mínima con la que se partirá para la consecución de este trabajo, siendo la actividad de plasmar los conocimientos de conceptos básicos.
- ¿Sobre qué bases de datos se realizará la exploración? Se debe realizar una investigación de las bases de datos genómicas existentes.
- ¿Es necesario el conocimiento de la estructura de las bases de datos genómicas? ¿Qué información es relevante? Se deberá realizar el análisis de la estructura de las bases de datos genómicas, para posterior a ello conocer que información es importante o relevante para este trabajo.
- ¿Cómo está vinculada la información que se presenta en las bases de datos genómicas con respecto a cada uno de los elementos del ECGH? Se realizará un mapeo de cada uno de los elementos del ECGH y la información que poseen las bases de datos genómicas.
- ¿Qué problemas se presentan en la exploración de las bases de datos genómicas? Se extraerá e indicará cada uno de los problemas encontrados en el momento de la exploración de las bases de datos genómicas.
- ¿Cómo se visualizará el resultado del mapeo de información? Se realizará una aplicación que permita la visualización de dichos resultados.
- ¿A qué conclusiones se ha llegado? Se expondrán las conclusiones a las que se ha llegado luego de la exploración de las bases de datos genómicas.

Ilustración 1 Metodología de investigación utilizada



Fuente: Propia

1. **Formulación del problema**: Se realizará una delimitación y planteamiento del problema para poder tener un punto central y claro al cual enfocar el proceso de la búsqueda de la solución.

2. **Conocimientos de conceptos básicos**: Se realizará la descripción de los conceptos tanto de Modelado conceptual así como de bases de datos genómicas para determinar relaciones existentes con respecto al ECGH.

3. **Investigación de bases de datos genómicas**: Se desarrollará una investigación de cada base de datos sobre la información que almacena y las distintas formas de acceso que proporciona.

4. **Análisis de la estructura de las bases de datos genómicas**: Descripción de la estructura de los datos que proporciona cada repositorio, en función de la forma de acceso.

5. **Mapeo de información**: Mapeo entre la información requerida por el ECGH y la información proporcionada por cada base de datos.

6. **Identificación de problemas**: Se realizará la identificación de los diferentes problemas que se encuentran en el momento de integran la información encontrada en cada una de las bases de datos genómicas al modelo conceptual genómico.

7. **Generación y presentación de resultados**: Posterior al mapeo e identificación de problemas se desarrollará una herramienta que permita visualizar los resultados del mapeo entre el ECGH y cada una de las bases de datos genómicas.

8. **Conclusiones**: Se sacarán las principales conclusiones encontradas de cada uno de los capítulos, así como los relacionados con los objetivos planteados en este trabajo.

1.5 ESTRUCTURA DEL TRABAJO

Este trabajo está compuesto por 6 capítulos estructurados de la siguiente manera:

- El **capítulo 1** introduce al lector en el contexto y alcance del trabajo, indicando la metodología y estructura del mismo.

- El **capítulo 2** presenta una introducción al Esquema Conceptual del Genoma Humano (ECGH), así como las características que este contiene. Adicionalmente se describen algunos elementos relevantes asociados con la biología molecular, especialmente con la importancia de la genética en el riesgo de sufrir determinadas enfermedades.

- El **capítulo 3** presenta una revisión de las diferentes bases de datos genómicas públicas en conjunto con sus tablas resumen en donde se pueden apreciar la información que manejan.

- El **capítulo 4** se presenta el mapeo realizado para cada uno de los elementos del modelo conceptual con la información presentada por las diferentes bases de datos genómicas investigadas. Se generará como resultado una aplicación para la presentación de resultados del mapeo de bases de datos genómicas y el ECGH.

- El **capítulo 5** contiene las diferentes conclusiones obtenidas del trabajo, ventajas, objetivos llevados a cabo, así como también trabajos futuros.

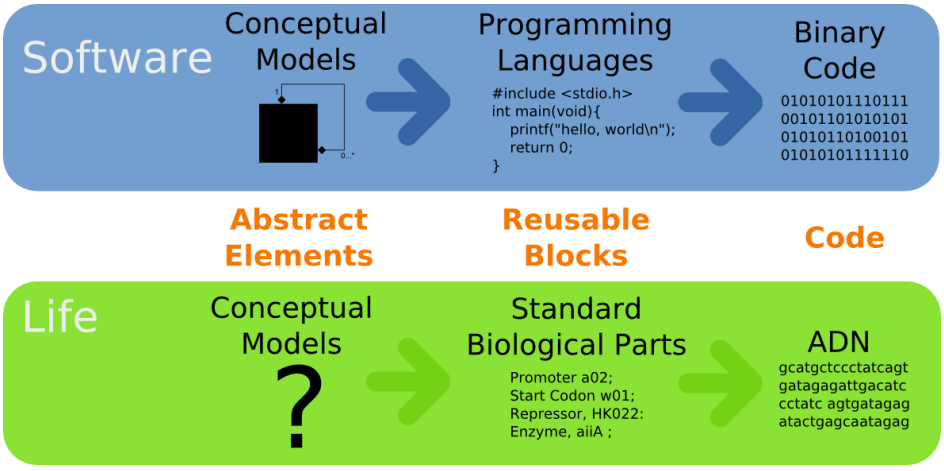
2 MODELO CONCEPTUAL

El Modelo Conceptual es utilizado en las diferentes áreas de trabajo ya que permite la abstracción de diferentes elementos para expresar su funcionamiento o su desempeño. Por su parte Mylopoulos en 1992 define al modelado conceptual como la actividad de describir formalmente algunos aspectos del mundo físico y social que nos rodea con la finalidad de comprender y comunicar. Adicionalmente menciona que este modelado está respaldado por diferentes estructuras, las cuales serán interpretadas por humanos y no por máquinas, promoviéndose una comprensión común entre el modelo y la realidad de donde fue tomada [14]. Thalheim precisa que el modelado se rige por su propósito, por ejemplo, la construcción de un sistema, simulación de situaciones del mundo real, construcción de teorías, explicación de fenómenos o documentación de un sistema existente. Siendo así el modelado también una actividad de ingeniería con pasos de ingeniería y resultados de ingeniería [15].

En sí el modelo conceptual es una herramienta que permite de manera versátil una descripción visual de las relaciones entre los elementos que intervienen en una tarea, actividad o proceso. Muchos autores llegan a considerar al modelo conceptual como un proceso abstracto para desarrollar una vista alternativa de la situación del problema, este modelo posteriormente permitirá que al volver al mundo real se pueda evaluar y efectivamente probar el funcionamiento del modelo creado [16]. Si bien no existe una descripción precisa donde se enmarque cada uno de los artefactos del modelado de manera rigurosa [17], la comunidad científica está en la posición de presentar mejoras o debatir sobre su elaboración [18].

Una de las aplicaciones del modelado conceptual es dentro del área informática en donde para la elaboración de software se plasma mediante modelos conceptuales el funcionamiento que debe tener, posteriormente el modelado se traduce a código en donde se realizan las tareas indicadas anteriormente en el modelado [19]. Se puede apreciar en la Ilustración 2 como es el cambio para software, partiendo de un modelo conceptual se pasa al lenguaje de programación y este a nivel máquina en secuencia de ceros y unos, de manera similar para modelar la vida se utiliza el modelado conceptual que contiene elementos extraídos de la realidad, los cuales en bloques reusables poseen los diferentes elementos propios de la cadena de ADN.

Ilustración 2 De los modelos conceptuales al código: una perspectiva SE y una perspectiva de comprensión de la vida.



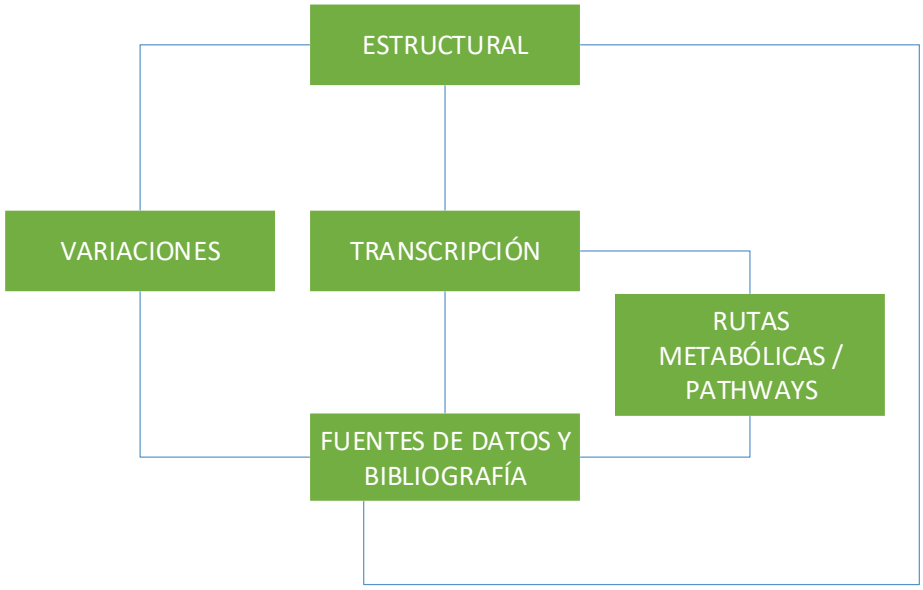
Fuente: Tomado del libro "Conceptual modeling perspectives" [19]

De la misma manera la aplicación del modelado conceptual es aplicado al ámbito médico, donde se consideran los diferentes procesos y se detallan las tareas y actividades que se desarrollan. Es aquí donde el modelado del genoma humano es una pieza fundamental que permite considerar las entidades involucradas y sus relaciones [20].

2.1 COMPONENTES BÁSICOS DEL MODELO CONCEPTUAL GENÓMICO

En base al Modelo conceptual expuesto por José Reyes (Ilustración 3), se puede observar 5 vistas que agrupan a cada uno de los elementos del ECGH [20]. Dicho ECGH ha sido desarrollado y propuesto de manera holística, por lo que en este trabajo se enfocará de la misma manera considerando todas las vistas. Se debe tener en cuenta que algunas de las vistas, dependiendo del área de investigación, pueden ser las más explotadas en el ámbito genético para reportes médicos o con fines netamente investigativos.

Ilustración 3 Vistas del Modelo Conceptual holístico



Fuente: Propia

Cada una de las diferentes vistas constan de un grupo de clases las cuales recogen los diferentes atributos propios de ellas, a continuación, se da una breve descripción de las mismas:

- ***Vista estructural***: La vista estructural presenta la información relacionada con la estructura del genoma asociado a una especie.

- ***Vista de transcripción***: La vista de transcripción básicamente contiene la información relacionada a la síntesis de proteínas.

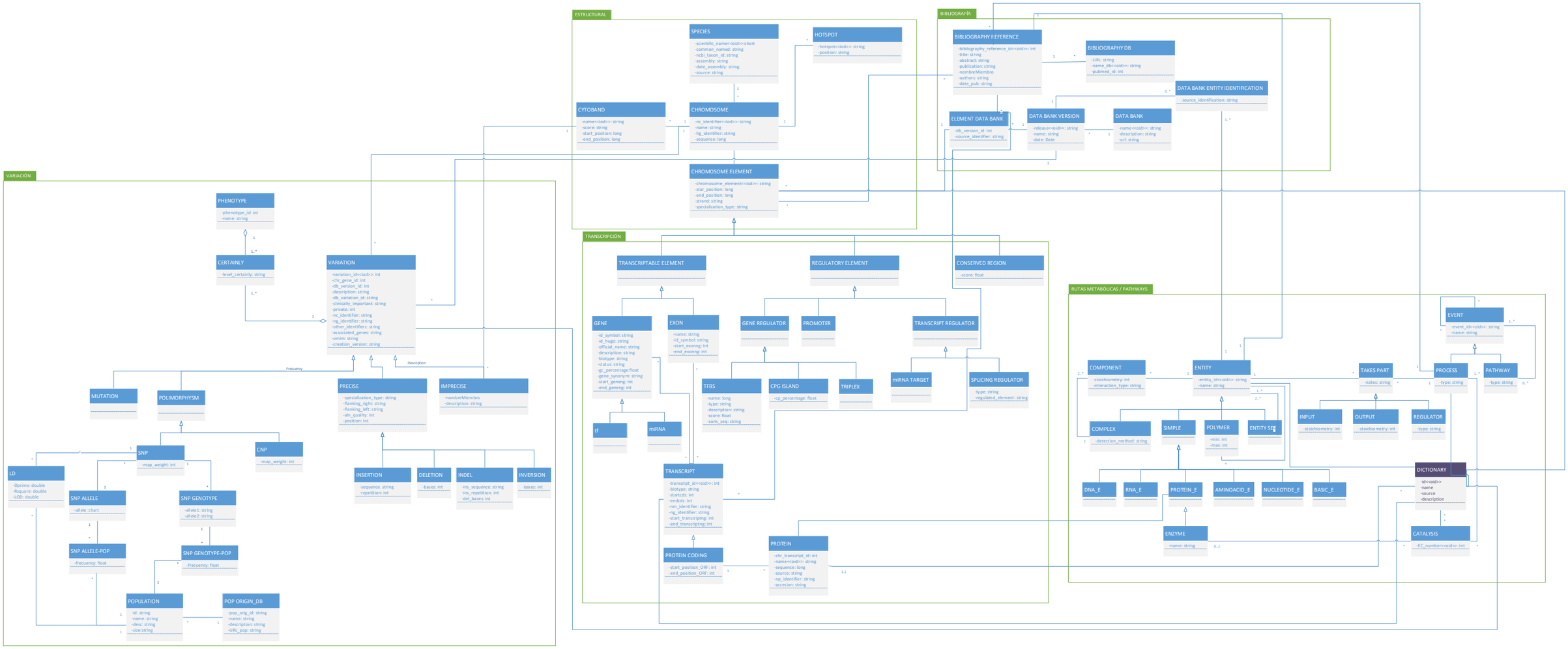
- ***Vista de variaciones***: La vista variación representa a cada una de las posibles variaciones que pueden presentarse en el ADN acorde a lo que hasta ahora la comunidad científica ha podido recabar.

- ***Vista de rutas metabólicas o pathways***: La vista de rutas metabólicas contiene la información sobre las reacciones químicas que suceden dentro de una célula.

- ***Vista de fuentes de datos y bibliografía***: La vista Bibliografía contiene la información relacionada con los repositorios que argumentan, dan validez científica y sobre todo que permiten la obtención de información de elementos cromosómicos, variaciones, poblaciones, proteínas, entre otras.

De manera más ampliada se presenta en la Ilustración 4 los componentes básicos del Modelo conceptual genómico, en los cuales se puede observar las clases que pertenecen a las diferentes vistas, así mismo se describe de manera extendida en el ANEXO 1 – ESQUEMA CONCEPTUAL DEL GENOMA HUMANO, todos los componentes para todas las vistas del ECGH.

Ilustración 4 Modelo conceptual a considerarse



Fuente: Tomado de la tesis doctoral de José Reyes [20]

3 BASES DE DATOS GENÓMICAS

En la actualidad el manejo de bases de datos se ha vuelto muy común, por lo que fácilmente se encuentra información recopilada en las diferentes áreas de trabajo. Es así como, en el ámbito genómico existe una gran variedad de bases de datos genómicas, es por ello que con la aparición de las NGS o también conocida como Next Generation Sequencing [21] (Secuenciación de segunda generación) es una de las impulsoras en que la información que se maneja sea de gran ayuda para reportes médicos, así como para investigaciones.

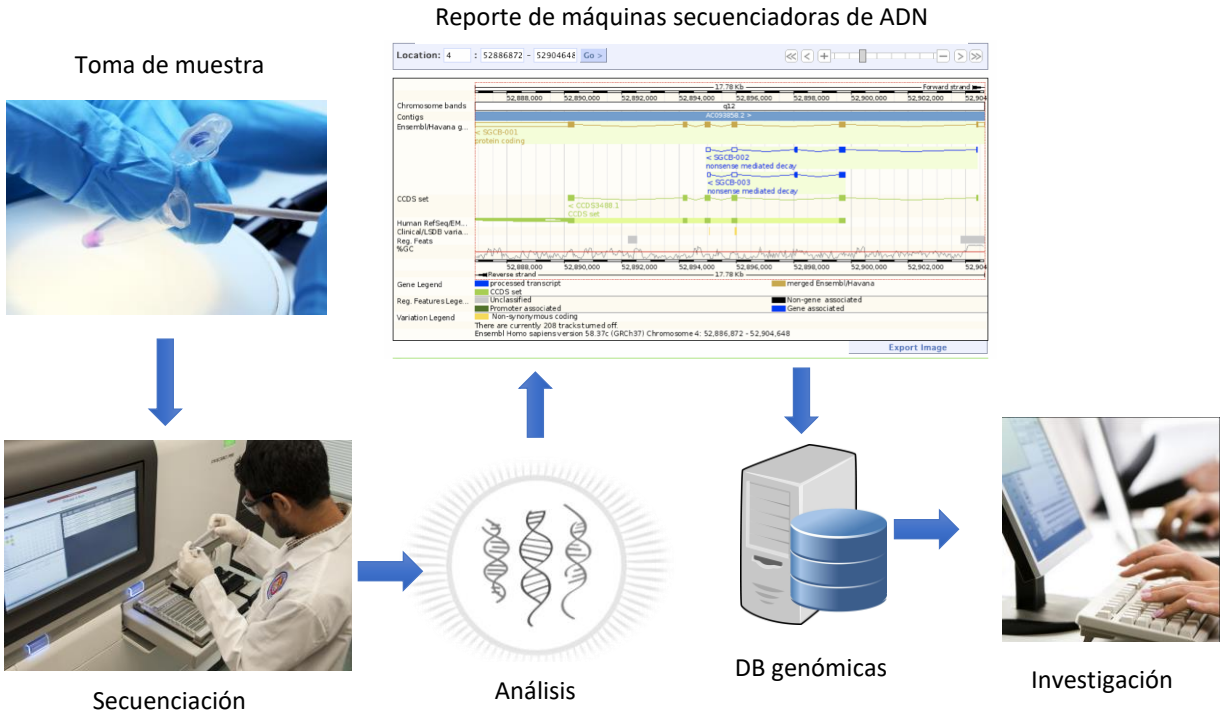
En esta sección se realizará una extracción de las diferentes bases de datos genómicas que aportan con diferentes atributos dentro del esquema conceptual del genoma humano. Para ello se han considerado las bases de datos a partir de las listados propuestos por Oxford Academic [22], National Center for Biotechnology Information (NCBI) [23], Human Genome Variation Society (HGVS) [24] y Health Sciencies Library System (HSLS) [25].

Se verificará la duplicidad de bases de datos genómicas que pudieren estar en más de un listado, obteniéndose como resultado un resumen de las bases de datos acompañadas de una breve descripción de cada base de datos, autor, fecha de creación, URL de acceso. Dicho resumen de bases de datos servirá para la sección 4 para mapear con el esquema conceptual del genoma humano tratado en la sección 2 MODELO CONCEPTUAL.

3.1 BASES DE DATOS GENÓMICAS

A medida que el tiempo avanza y con este los diferentes progresos en el ámbito genético, se ha podido avizorar que existe información genómica en múltiples fuentes por lo que debe ser canalizada y clasificada. Una base de datos genómica, o conocida también como base de datos del genoma, es el repositorio en el cual se puede realizar el mapeo genómico resultante de lo encontrado en las pruebas de laboratorio [26]. Actualmente las bases de datos genómicas son herramientas de vital importancia ya que permiten realizar investigaciones y estudios en base al contenido o enfoque que se ha dado dentro del contenido de estas [27]. Los expertos en el ámbito biológico son los principales beneficiarios de dicha información, puesto que bajo la estructura con la que fue concebida la base de datos permite comprender diferentes fenómenos biológicos.

Ilustración 5 Origen de bases de datos genómicas



Fuente: Propia

Las bases de datos genómicas fueron concebidas a partir de las bases de datos relacionales enfocadas al ámbito biológico en las cuales el diseño, desarrollo y gestión a largo plazo han llegado a formar el punto central en el ámbito de la bioinformática [28]. El contenido de dichas bases de datos es el resultado de la secuenciación de ADN a partir de una muestra de un ser vivo (Ver Ilustración 5). Posteriormente las máquinas de secuenciación que permiten obtener el reporte genómico, el cual es procesado por los expertos (genetistas en su gran mayoría). Los expertos que tratan la información genómica en base a su técnica, área, especie de estudio, entre otros son quienes generan las bases de datos genómicas con diferentes tipos de información entre ellas secuencias de nucleótidos, expresiones génicas, genes de resistencia a antibióticos, taxonomía, genomas, mutaciones, variaciones genéticas, estructura secundaria de proteínas, familias, dominios, índice de publicaciones de artículos científicos, rutas de interacción, reacción de proteínas y enzimas, entre otras [22] [23] [25] [29].

En muchas ocasiones, cuando se requiere obtener las diferentes fuentes de información, por ejemplo la nomenclatura de un gen o artículos de una variación que han sido generados como sustento de resultados o investigaciones, el resultado que tenemos simplemente es limitarse a las fuentes de información más conocidos o más utilizados en el medio. La tarea se torna tediosa y muy extensa en ocasiones en las que el investigador desea expandir la búsqueda de fuentes, por lo que en su defecto el primer paso es llevarlo a un buscador en internet, en donde no se dispone un resumen o sumario donde se pueda tener todos estos resultados. Cabe indicar que no siempre se tiene el mismo número de fuentes, algunas han sido dadas de baja y otras quizás pudieron surgir. Es así que se requiere de un lugar donde se pueda obtener un acercamiento a una síntesis de fuentes donde recurrir y que mejor si se lo toma en base al ECGH.

Con el paso del tiempo se han generado diversas bases de datos que contienen información genómica, estas bases de datos adicionalmente forman parte de listados en los cuales el público puede acceder de manera libre o privada (de pago o algún tipo de membresía), tal es el caso de los listados proporcionados por Oxford Academic [22], National Center for Biotechnology Information (NCBI) [23], Human Genome Variation Society (HGVS) [24] y Health Sciences Library System (HSLS) [25].

Los listados surgen a partir de iniciativas de grupos de investigación, dichos grupos han intentado agrupar diversas bases de datos genómicas que han ido apareciendo a lo largo del tiempo. Sin embargo, he aquí donde aparecen debilidades:

- Las bases que se presentan en los distintos listados no son las mismas. Cada listado presenta bases de datos que difieren en su contenido.
- Los elementos que se indican en cada base de datos son diferentes, en algunas bases de datos puede indicarse más a detalle la información o presentarse más atributos al respecto. Esto se debe a que queda a criterio de los autores o administradores del grupo de investigación que maneja la información.

3.1.1 OXFORD ACADEMIC

La Oxford University Press es un departamento de la Universidad de Oxford, el cual tiene una sección llamada NAR (Nucleic Acid Research) destinada para la recepción y publicación de información científica orientada a los resultados de investigaciones de vanguardia sobre los aspectos físicos, químicos, bioquímicos y biológicos de los ácidos nucleicos y las proteínas implicadas en el metabolismo y/o las interacciones de los ácidos nucleicos [30].

Oxford University Press y los Editores de NAR lanzaron una iniciativa de acceso abierto para NAR en 2005. Esto significa que ya no es necesario tener una suscripción en orden para leer el contenido actual de NAR en línea [30].

Actualmente NAR presenta un total de 1610 bases de datos, las cuales presentan una categorización como indica la Tabla 1, permitiendo tener un gran número de datos para cada categoría.

Tabla 1 Categorías de DB en NAR

CATEGORÍA	DESCRIPCIÓN
NUCLEOTIDE SEQUENCE DATABASES	Contiene información de secuencias de nucleótidos, ADN codificado y no codificado, estructura de genes, intrones y exones, sitios de empalme, factores de transcripción.
RNA SEQUENCE DATABASES	Se refiere a secuencias de RNA (Ácido Ribonucleico).
PROTEIN SEQUENCE DATABASES	Posee información de secuencias generales, propiedades de las proteínas, localización de proteínas, clasificación de proteínas, familias de proteínas.
STRUCTURE DATABASES	Tiene información de moléculas pequeñas, hidratos de carbono, estructuras de ácidos nucleicos y estructuras proteicas.

CATEGORÍA	DESCRIPCIÓN
GENOMICS DATABASES (NON-VERTEBRATE)	Presenta información de términos de anotación del genoma, ontologías y nomenclaturas, taxonomía, genomas virales, genomas procarióticos, genomas de eucariotas unicelulares, genomas fúngicos, genomas de invertebrados y genómica comparativa.
METABOLIC AND SIGNALING PATHWAYS	Indica información de enzimas, nomenclatura enzimática, rutas metabólicas, interacciones entre proteínas.
HUMAN AND OTHER VERTEBRATE GENOMES	Posee información de genoma de invertebrados, genómica comparativa, mapas y visores del genoma humano.
MICROARRAY DATA AND OTHER GENE EXPRESSION DATABASES	Contiene información general de genética humana, polimorfismo, datos específicos de genes, sistemas o enfermedades.
PROTEOMICS RESOURCES	Presenta información de recursos de proteómica.
OTHER MOLECULAR BIOLOGY DATABASES	Incluye base de datos de literatura, drogas, sondas moleculares.
ORGANELLE DATABASES	Asocia información de genes y proteínas mitocondriales.
PLANT DATABASES	Indica información de bases de datos de plantas, Arabidopsis thaliana, arroz entre otras.
IMMUNOLOGICAL DATABASES	Contiene información referente a inmunología.
CELL BIOLOGY	Presenta información de biología celular.

Fuente: Tomado de NAR [31]

3.1.2 NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

El Centro Nacional de Información Biotecnológica (NCBI) inició con la finalidad de identificar los registros de GenBank que en esas fechas donde se recopilaba la secuencia del genoma de manera aceptada por un organismo particular. No obstante, en la década de los 90's cuando se obtuvo el genoma del VIH-1 no se encontraba completo, por lo que muchos investigadores se organizaron para tener en común un genoma con el cual podrían trabajar sin problemas, pero lamentablemente ese genoma no estuvo depositado en GenBank. Posteriormente el equipo de NCBI trabajó en conjunto con dichos investigadores en la elaboración de un libro sobre retrovirus, donde se vinculaba nuevamente a GenBank como un repositorio de genomas que se utilizaban para la investigación [32]. NCBI asume la responsabilidad de la base de datos de secuencias de ADN de GenBank en octubre de 1992.

A partir de la consolidación de varios genomas de los investigadores, NCBI permitió que se pueda de manera confiable y robusta la descarga, visualización o análisis de diferentes recursos, herramientas y códigos compartidos por la comunidad científica. Actualmente se puede encontrar material como genes, genomas, secuencias de referencia, anotaciones de genomas eucarióticos o procarióticos, variaciones, entre otros. Algunos recursos reflejan los principios organizadores naturales de datos genómicos y apoyan el acceso desde una perspectiva centrada en genes u organismos, mientras que otros se desarrollaron en respuesta a un brote de enfermedad particular o en colaboración, comentarios o solicitudes de la comunidad. A lo largo

del camino también se desarrolló un conjunto de plataformas de visualización que incluyen Visores como Map Viewer, visor de secuencia gráfica y una interfaz gráfica de usuario Genome Workbench que puede ser descargado y es multiplataforma [32] [33] [34] [35].

Actualmente NCBI contiene 41 bases de datos las cuales han sido categorizadas como presenta la Tabla 2, siendo así que abarca diferentes ámbitos en la que los investigadores de manera pública pueden acceder a los contenidos. El personal de NCBI tiene capacitación avanzada en biología molecular, que son quienes construyen la base de datos a partir de secuencias enviadas por laboratorios individuales y mediante el intercambio de datos con las bases de datos de secuencias de nucleótidos internacionales, European Molecular Biology Laboratory (EMBL) Base de datos de ADN de Japón (DDBJ) [32].

Tabla 2 Categorías de DB en NCBI

CATEGORÍA	DESCRIPCIÓN
LITERATURE	Orientada al contenido de libros, reportes o resúmenes que indican resultados vinculados con el genoma.
GENES	Posee información sobre secuencias etiquetadas, transcripciones, expresiones génicas, genes homólogos, estudios filogenéticos y de población.
GENETICS	Se relaciona a variaciones humanas, genotipos, fenotipos, variación estructural del genoma, registros de pruebas genéticas, herencia mendeliana en el hombre.
PROTEINS	Posee información de dominios de proteínas conservadas, secuencias de proteínas, estructuras biomoleculares.
GENOMES	Contiene información de secuencias de AND y ARN, secuencias genómicas, taxonomía, nomenclaturas.
CHEMICALS	Presenta información de rutas metabólicas, estudios de detección de bioactividad, información química con estructuras, información química.

Fuente: Tomado del sitio web de NCBI [23]

3.1.3 HUMAN GENOME VARIATION SOCIETY

La Sociedad de Variación del Genoma Humano (HGVS) es un afiliado de la Federación Internacional de Sociedades de Genética Humana (IFHGS) y también de la Organización del Genoma Humano (HUGO). Tiene como objetivo fomentar el descubrimiento y la caracterización de las variaciones genómicas, así como también la distribución de la población y las asociaciones fenotípicas. Además, promueve la recopilación, documentación y distribución gratuita de la información de variación genómica y las variaciones clínicas asociadas. Otro de sus objetivos es fomentar el desarrollo de la metodología y herramientas necesarias para ayudar en la investigación [36].

La misión de HGVS es mejorar la salud humana a través de la identificación y caracterización de los cambios en el genoma que conducen a la susceptibilidad de la enfermedad. De esta manera HGVS permite cotejar la información genómica necesaria para el diagnóstico molecular, la investigación sobre los mecanismos básicos y el diseño de los tratamientos de las dolencias humanas [36].

Inicialmente el grupo de HGVS propuso diferentes descripciones para las variantes de secuencias en las secuencias de ADN y proteínas (mutaciones, polimorfismos), esto fue presentado en dos artículos publicados en 1993, desde ahí han existido modificaciones que han permitido definir detalles que en un inicio no fueron considerados o en su defecto abarcar nuevos descubrimientos. Es así que a la fecha han existido 10 actualizaciones sobre la versión original, dando como resultado las recomendaciones del año 2016 para la secuencia de variaciones que permiten a los diversos grupos de investigación incorporar sus bases de datos en base a estos parámetros [37] [38].

Actualmente HGVS ha agrupado bases de datos genómicas bajo las categorías indicadas en la Tabla 3. Esta categorización ayuda a los investigadores a encontrar el material necesario para sus trabajos, así como las herramientas informáticas que se requieran [24].

Tabla 3 Categoría de DB de HGVS

CATEGORÍA	DESCRIPCIÓN
LOCUS SPECIFIC MUTATION DATABASES	Contiene información de las bases de datos de mutaciones específicas de Locus, es decir en base a una posición fija en un cromosoma.
DISEASE CENTERED CENTRAL MUTATION DATABASES	Bases de datos de mutaciones centrales enfocadas en la enfermedad, por ejemplo asma, Parkinson, desordenes de retina, síndrome de Werner.
CENTRAL MUTATION & SNP DATABASES	Agrupa bases de datos sobre polimorfismo, biología molecular, variaciones genómicas, mutaciones génicas, SNP, farmacogenética, farmacogenómica.
NATIONAL & ETHNIC MUTATION DATABASES	Contiene información de variaciones genéticas de diferentes poblaciones, por ejemplo Corea, China, Singapur, Iraní, Israelí, Árabe, Chipriota
MITOCHONDRIAL MUTATION DATABASES	Presenta información de mutaciones mitocondriales, herencia mendeliana.
CHROMOSOMAL VARIATION DATABASES	Se encuentran colecciones de anomalías cromosómicas, variación cromosómica en el hombre, archivos de polimorfismos de inserción de retrotransposones, aberraciones cromosómicas en cáncer.
OTHER MUTATION DATABASES	Información de sistemas para receptores nucleares, trastornos hereditarios, mutaciones de filamentos humanos intermedios.
CLINICAL & PATIENT ASPECTS DATABASES	Contiene información sobre Alzheimer, glucosa, diabetes, cáncer de páncreas.
NON HUMAN MUTATION DATABASES	Información de variaciones en especies no humanas, entre ellas insectos, ratones, peces, bacterias, ribosomas.
ARTIFICIAL MUTATIONS ONLY	Contiene información referente a enzimas, mutaciones génicas de mamíferos, receptores acoplados a proteínas, mutaciones del canal Ion.
OTHER RELATED DATABASES	Presenta un conjunto de diversas bases de datos entre ellas con información de alelos, macromoléculas, genomas australianos, segregaciones cromosómicas, secuencia de nucleótidos, tumores, entre otros.
EDUCATION RESOURCES FOR TEACHERS & STUDENTS	Es un compendio de recursos los cuales se enfocan en temas biológicos, siendo algunos recursos interactivos.

Fuente: Tomado del sitio web de HGVS [24]

3.1.4 HEALTH SCIENCES LIBRARY SYSTEM

El Sistema de Bibliotecas de Ciencias de la Salud (HSLS) de la Universidad de Pittsburgh ofrece una amplia gama de servicios de información, oportunidades educativas y recursos en formato impreso y electrónico para profesores, estudiantes e investigadores en las escuelas de ciencias de la salud (Medicina, Medicina Dental, Farmacia, Enfermería, Ciencias de la Salud y Rehabilitación, y Salud Pública). Es así como en su sitio web presenta un listado de bases de datos relacionados con el ámbito genómico que son de acceso público [39].

En mayo de 2011, HSLS recibió un contrato de cinco años de la Biblioteca Nacional de Medicina para servir como Biblioteca Médica Regional (RML) para la Región del Atlántico Medio de la Red Nacional de Bibliotecas de Medicina (NN/LM-MAR). La misión de MAR es apoyar los esfuerzos de la Biblioteca Nacional de Medicina para proporcionar a todos los profesionales de la salud de EE.UU. acceso equitativo a la información biomédica y mejorar el acceso público a la información para que puedan tomar decisiones informadas sobre su salud. El MAR es una de las ocho regiones de la Red Nacional de Bibliotecas de Medicina (NN / LM) e incluye los estados de Delaware, Nueva Jersey y Nueva York y Pensilvania [40].

HSLS apoya la enseñanza, investigación y atención clínica, enfocado para los estudiantes y personal de la facultad de ciencias de la salud de la Universidad de Pittsburgh, así como para los residentes y becarios del Centro Médico de la Universidad de Pittsburgh [41]. Actualmente los repositorios de HSLS posee información de investigaciones sobre temas de salud y biomédicos, revisiones sistemáticas y otras búsquedas avanzadas de literatura, gestión de datos de investigación, recursos de biología molecular y herramientas de software, entre otros [42].

Actualmente HSLS maneja 2458 bases de datos, las cuales se agrupan en las categorías indicadas en la Tabla 4.

Tabla 4 Categorías de BD de HSLS

CATEGORÍA	DESCRIPCIÓN
DNA SEQUENCE DATABASES AND ANALYSIS TOOLS	Presenta información de secuencias de AND, secuencias de nucleótidos, oligones.
ENZYMES AND PATHWAYS	Contiene información de enzimas, rutas metabólicas, interacción entre proteínas, señalización en rutas metabólicas.
GENE MUTATIONS, GENETIC VARIATIONS AND DISEASES	Posee información sobre mutaciones, polimorfismos, enfermedades y proteínas.
GENOMICS DATABASES AND ANALYSIS TOOLS	Asocia información de comparativas genómicas, herramientas y bases de datos genómicas, anotaciones, ontologías, análisis de secuencia genómicas, mapas, visores, datos genómicos de organismos vertebrados no humanos.
IMMUNOLOGICAL DATABASES AND TOOLS	Contiene diferentes bases de datos relacionados con inmunología, así como herramientas.
MICROARRAY, SAGE, AND OTHER GENE EXPRESSION	Presenta información de diseño de microarray, sondas, herramientas de análisis de datos de expresión génica.
ORGANELLE DATABASES	Es un compendio de diferentes bases de datos de organelos.

OTHER DATABASES AND TOOLS (LITERATURE MINING, LAB PROTOCOLS, MEDICAL TOPICS, AND OTHERS)	Maneja bases de datos de literatura, protocolos de laboratorio, diseño de medicamentos y herramientas no clasificadas.
PLANT DATABASES	Contiene información de plantas, tales como: Arabidopsis thaliana, arroz, entre otras.
PROTEIN SEQUENCE DATABASES AND ANALYSIS TOOLS	Posee información de secuencias de proteínas generales, alineación de similitud de secuencias, familias de proteínas, dominios proteicos, filogenia, secuencias de proteínas, anotaciones.
PROTEOMICS RESOURCES	Recoge información de diferentes bases de datos referente a proteínas.
RNA DATABASES AND ANALYSIS TOOLS	Indica información de interferencia, secuencias, estructuras y funciones de ARN.
STRUCTURE DATABASES AND ANALYSIS TOOLS	Presenta información referente a hidratos de carbono, estructuras de ARN, proteínas y pequeñas moléculas.

Fuente: Tomado del sitio web de HSLS [29].

Como se ha observado Oxford, HSLS, NCBI, HGVS presentan distintas clasificaciones o categorizaciones para las bases de datos que constan en sus listados, no obstante, ya se visualiza que no agrupan todos los listados a todos los elementos en cuanto a categorías. Inicialmente los trabajos o proyectos de los que se derivan los datos de origen y resultados son publicados, pero bajo criterios de los autores y es aquí donde la diversificación de las bases de datos se genera.

Claramente se puede apreciar que ningún listado es completo, debido a esta razón es que se ha agrupado las bases de datos de los 4 listados, de esta manera se abarcará y revisará más contenido. De inicio se tendrían bases de datos que estén orientadas a elementos de ADN de diferentes especies, secuencias cromosómicas, proteínas, variaciones, enzimas, rutas metabólicas, enfermedades, mutaciones, genes, inmunología, células, RNA, entre otros.

Es importante tener en cuenta que algunos listados constantemente van actualizándose, esto para el caso de agregar nuevas bases de datos genómicas. Pero así mismo algunas bases de datos puedan no estar contenidas en ellos. Esto puede ser por el hecho que no se han dado a conocer o utilizado en alguna investigación hasta el momento que los haya referenciado.

Considerando estos aspectos el desarrollo de este trabajo pretende abarcar un segmento más grande al que ha llegado un sitio con un listado de bases de datos, así mismo contener más bases de datos genómicas, en las cuales se pueda diversificar la información como resultado. Esto permitirá al investigador que pueda enlazar datos desde diferentes fuentes al mismo tiempo que su investigación tenga más argumentos para probar sus hipótesis.

4 MAPA DE BASE DE DATOS GENÓMICAS SOBRE EL ESQUEMA CONCEPTUAL DEL GENOMA HUMANO

En esta sección se considera la lista depurada de las diferentes bases de datos genómicas que se han encontrado y se contrastará con el esquema conceptual del genoma humano (ECGH). Como resultado del mapeo se podrá obtener trazabilidad de entre los diferentes atributos de las clases consideradas en el ECGH y las diferentes fuentes con su correspondiente en la base de datos que disponga de esta información. Es de vital importancia en esta sección primeramente fijar diferentes criterios de selección de las bases de datos ya que en un principio se puede considerar todas las que se presenten en la web, pero no todas pueden ser categorizadas u orientadas al trabajo que se pretende obtener como resultado, por esta razón como primer paso se establecerán criterios que conlleven a una fácil alineación de los conceptos a obtenerse. Así mismo, se podrá desde esta sección se obtendrá la información necesaria para la generación de la aplicación de presentación de resultados de la exploración de bases de datos genómicas. Por otro lado, se identificarán los problemas que se han encontrado o pudiesen existir para todo este proceso.

4.1 CRITERIOS DE SELECCIÓN

Para la selección de las bases de datos genómicas se han considerado los listados ofertados por NCBI, Oxford, HVGS y HSLS, indicados en la sección 3.1. Tomando como referencia la sección 2, se han definido criterios de cumplimiento que permiten recabar información que se utiliza en el Esquema Conceptual del Genoma Humano. A continuación, se detallan los criterios que deben cumplir:

1. Considerando las diferentes vistas que plantea el esquema conceptual de la sección 2, las bases de datos deben poseer información de:

    a. ***Genes***: Bases de datos que muestre información sobre genes, ya sea nomenclaturas, descripciones, expresiones, entre otros.
    b. ***Genomas***: Serán consideradas las bases de datos en las que se disponga de información de genomas de diversas especies o poblaciones, la cual esté acorde al ECGH.
    c. ***Especies***: Se consideran para la selección de las bases de datos que dispongan de información sobre especies, en este caso al tratarse del genoma humano se filtrará todas las bases de datos que presenten información humana. No obstante, las bases de datos que indiquen información de otras especies se presentarán en un listado indicando la especie a la que corresponden.
    d. ***Proteínas***: Las bases de datos que muestren información de proteínas, dentro de la cual se pueda ver descripción, nomenclaturas, ubicación, entre otros.
    e. ***Químicos***: Se considerarán las bases de datos que indiquen información sobre químicos que estén relacionados con alteraciones o variaciones correspondientes al genoma.
    f. ***Pathways***: La información que se enmarque en las rutas metabólicas serán también consideradas como un criterio de selección en las bases de datos.

g. ***Variaciones***: Se considera la información de las bases de datos que esté relacionado con las diferentes variaciones presentes en humanos.
h. ***Poblaciones***: Las bases de datos genómicas que contienen información sobre las poblaciones en las que se ha realizado los estudios o de donde proviene la información.
i. ***Bibliografía***: Todas las bases de datos que dispongan información de referencias científicas sobre los estudios realizados.

2. Credibilidad de las fuentes: Las fuentes han de ser consideradas las páginas publicadas o referenciadas por entidades ya sean educativas o dedicadas al ámbito de la salud. Así mismo se indicará si la información de la base de datos ha sido revisada o existe un autor (personas/personas/organización) que avale la veracidad de la información publicada.

4.2 SELECCIÓN DE BASES DE DATOS GENÓMICAS

Para la selección de las bases de datos genómicas primeramente se han considerado las presentadas en 4 colecciones, la primera las bases que están asociadas a NBCI [23], segunda la colección online de recursos bioinformáticos de la Universidad de Pittsburgh [25], la cual presenta en línea un conjunto de anotaciones y enlaces para bases de datos bioinformáticos así como herramientas de software. Así mismo también se han considerado la colección presentada por la Academia de Oxford [22] en la que se encuentran artículos científicos sobre las diferentes investigaciones tanto a nivel biológico, así como también químico, físico y bioquímico, su acceso es totalmente abierto lo que genera gran facilidad en el acceso al contenido.

Es importante mencionar que a pesar de tomar toda la información en julio de 2018 los sitios web presentan un copyright del año 2014. Adicionalmente se consideró el listado propuesto por la Sociedad de variación del genoma humano (HGVS) [24], cuyo objetivo es fomentar el descubrimiento y la caracterización de las variaciones genómicas, incluida la distribución de la población y las asociaciones fenotípicas.

4.2.1 EXTRACCIÓN DE LISTADOS DE BASES DE DATOS

Para el caso de la generación de un listado único en base a las fuentes consultadas, primeramente, se revisó los sitios web en los cuales simplemente se presentan mediante un HTML el texto de las bases de datos. Ninguno de los sitios permite la descarga de archivos csv o txt por ejemplo en donde se disponga ya de los nombres, descripción, URL, que es información útil para conformar una base de datos inicial, por este motivo se torna lenta la extracción de información al tener que copiar cada uno de estos atributos en una hoja de cálculo o ir ingresando manualmente en una base de datos.

Como una forma de agilizar este proceso se generaron diferentes scripts en donde se realizó filtrados de información para la extracción de la información. Los scripts que se generaron básicamente permiten mediante Python obtener atributos básicos de cada base de datos:

1. ***Oxford*** (Ver Ilustración 6): se guardó la página web [30] seleccionando el listado alfabético, de esta manera se consideraron todas las bases de datos que el repositorio tenía. Los datos importantes para extraer son:
    1. Nombre de la base de datos, solo se extrae directamente el texto.

2. Autores de la base de datos, se comprueba que solo se extraiga el nombre de las personas sin etiquetas HTML.
3. Descripción de la base de datos, se extrae directamente el texto.
4. URL de la base de datos, se extrae el texto a partir del código HTML en donde se indica el enlace de la base de datos. En algunos casos se presenta 2 direcciones web por lo que inicialmente se deja intacto para en el momento de ir verificando cada una de las bases de datos se pueda considerar cuál de ellas es la válida.

Ilustración 6 Extracción de datos de Oxford Journal

Selección inicial: Listado general de DBs.

Fuente: Propia

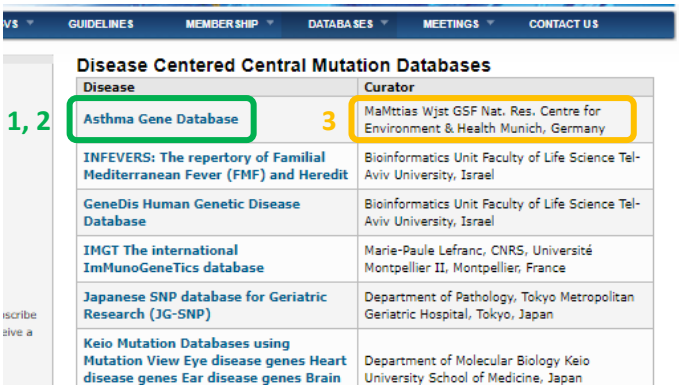
2. ***HVGS*** (Ver Ilustración 7): el primer acceso al listado es simplemente una categorización que HGVS propone, por lo que en su contenido no está indicado el nombre de las bases de datos, por lo que accedemos manualmente a cada una de las categorías (Ver Ilustración 8). El contenido de cada una de las páginas permite extraer:
    1. Nombre de la base de datos, extrayendo directamente el texto.
    2. URL de la base de datos, se verifican las etiquetas HTML y extrae el vínculo.
    3. Creadores de la base de datos, directamente tomando el texto proporcionado.

Ilustración 7 Revisión de HGVS

Categorías propuestas por HGVS

Fuente: Propia

Ilustración 8 Extracción de datos de HGVS



Fuente: Propia

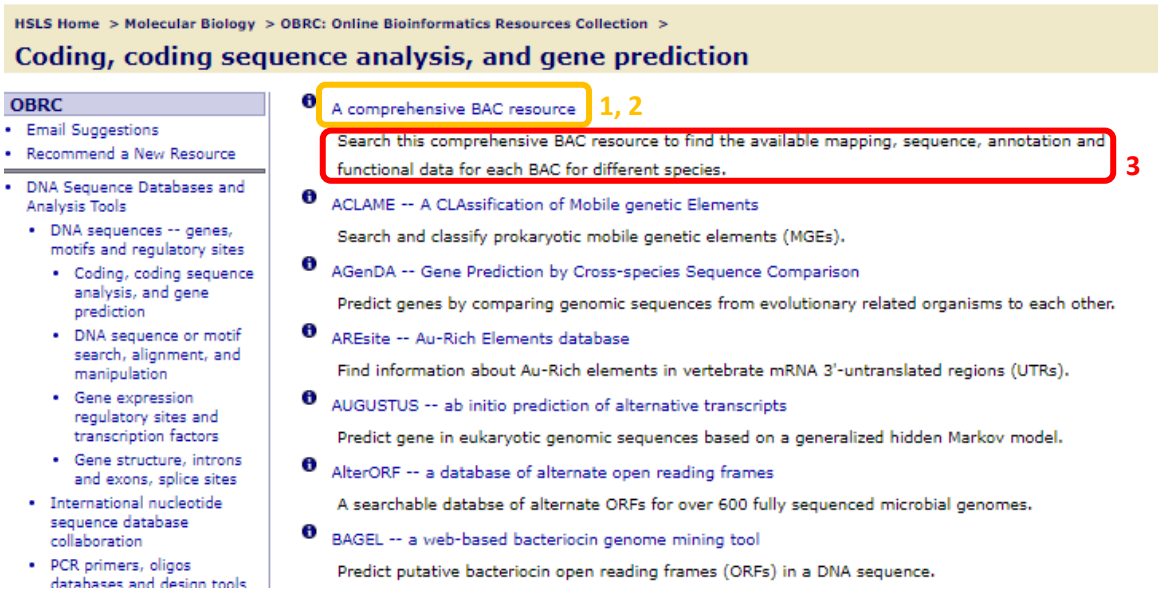
3. ***HSLS*** (Ver Ilustración 9): se guardó la página HTML y como en el caso de HGVS el contenido estuvo redireccionado a otras páginas web por lo que se accede a cada una de las categorías o subcategorías si es el caso y posteriormente (Ver Ilustración 10) a extraer los valores de cada base de datos.
    1. Nombre de la base de datos, extracción solo del texto.
    2. URL de la base de datos, se verifican las etiquetas HTML y extrae el vínculo.
    3. Extracción del texto que contiene la descripción.

Ilustración 9 Extracción de datos de HSLS

Categorías propuestas por HSLS

Fuente: Propia

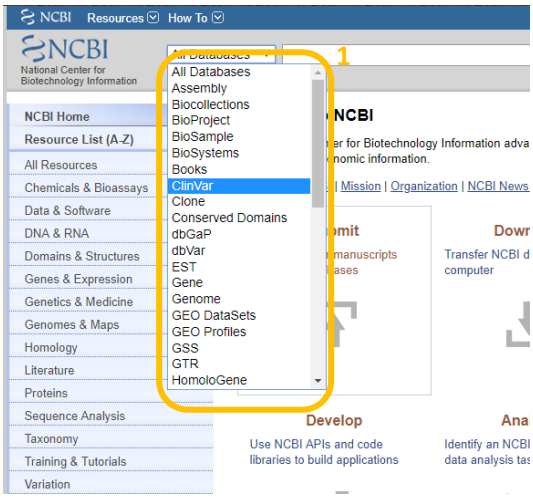
Ilustración 10 Selección de cada una de las categorías de HSLS



Fuente: Propia

4. Para el caso de NCBI presenta este listado dentro de un cuadro de selección como se observa en la Ilustración 11 donde está enmarcado en amarillo, por lo que se copia manualmente la información al ser solamente 43 datos.

Ilustración 11 Extracción de datos de NCBI



Fuente: Propia

Finalmente se obtuvieron 4 archivos correspondiente a cada uno de los listados fuente.

4.2.2 TECNOLOGÍAS EMPLEADAS

Para la realización de los scripts en la sección de limpieza de base de datos se ha seleccionado como IDE a Python 3 bajo su entorno de trabajo Jupyter Notebook dentro de Anaconda 3. Se han utilizado las siguientes librerías:

- ***io:*** Esta librería permite realizar la gestión de archivos planos [43].
- ***csv:*** La librería csv permite gestionar los archivos csv [44].
- ***httplib:*** Este módulo define las clases que se implementan al lado del cliente de los protocolos HTTP y HTTPS [45].
- ***urlparse:*** Este módulo define una interfaz estándar para dividir las cadenas de URL en componentes (esquema de direcciones, ubicación de red, ruta, etc.), para volver a combinar los componentes en una cadena de URL y para convertir una "URL relativa" en URL absoluta con una "URL base" [46].

4.2.3 LIMPIEZA DE BASES DE DATOS

En esta fase de diferentes puntos importantes para obtener un listado final de bases de datos genómicas mismo que nos servirá para realizar el mapeo entre cada una de las bases de datos y el modelo conceptual del genoma humano:

1. Se verificó duplicidad en nombres de cada una de las bases de datos que constan en el listado general, por lo que es de vital importancia ya que en el momento de acceder a las mismas se evitará realizar doble trabajo en dicha actividad. Cabe indicar que en algunas ocasiones se indica la base de datos con un acrónimo por nombre y en otros casos está con un nombre extendido por lo que este caso también es considerado. De esta manera inicialmente se presentaron 4552 bases de datos (Ver Tabla 5).

Tabla 5 Preselección de Bases de datos genómicas

FUENTE	NO.
NCBI	43
OXFORD	1610
HSLS	2401
HGVS	498
TOTAL	4552

Fuente: Propia

2. Se realizó una revisión de las URLs y se eliminó las que presentaban duplicidad, esto se debe a que en varias ocasiones las bases de datos genómicas presentaban diferentes nombres, pero sus enlaces apuntaban al mismo sitio.
3. Eliminación de caracteres especiales, y sintaxis no válidas.
4. Mediante la utilización de un script se determinó qué bases de datos tienen acceso (Carga a un sitio web) y las que no (No se carga ningún sitio web), esta determinación se hace en base al código de acceso que el script devolvió (Ver Tabla 6). En base a los códigos obtenidos, se seleccionaron las bases de datos que obtuvieron código 200, puesto que son las que no tienen ningún problema en acceso y conexión. Para el caso de las bases que obtuvieron códigos 400 (errores de cliente, la página está disponible pero el recurso no) y 500 (errores de servidor) fueron descartadas.

Tabla 6 Códigos de respuesta en comprobación de sitios web

CÓDIGO	DEFINICIÓN
200	Carga un sitio web sin problemas.
400	Bad Request (La petición contiene un error).
401	UnAutorizhed Se puede autenticar, pero no en el momento no puede ser servida la solicitud o ha fallado.
403	Petición bien realizada pero el servidor no responde por falta de permisos del usuario.
404	La página está disponible pero el recurso no.
410	El recurso se ha quitado permanentemente, no está ni estará disponible.
412	La petición no se cumple por todas las condiciones impuestas por el cliente.
500	El servidor no puede dar respuesta por caída del sitio.
502	Existe un servidor de enlace Gateway, el cual no está disponible.
503	Servidor congestionado o en mantenimiento.
504	Timeout en la respuesta por demora del servidor.

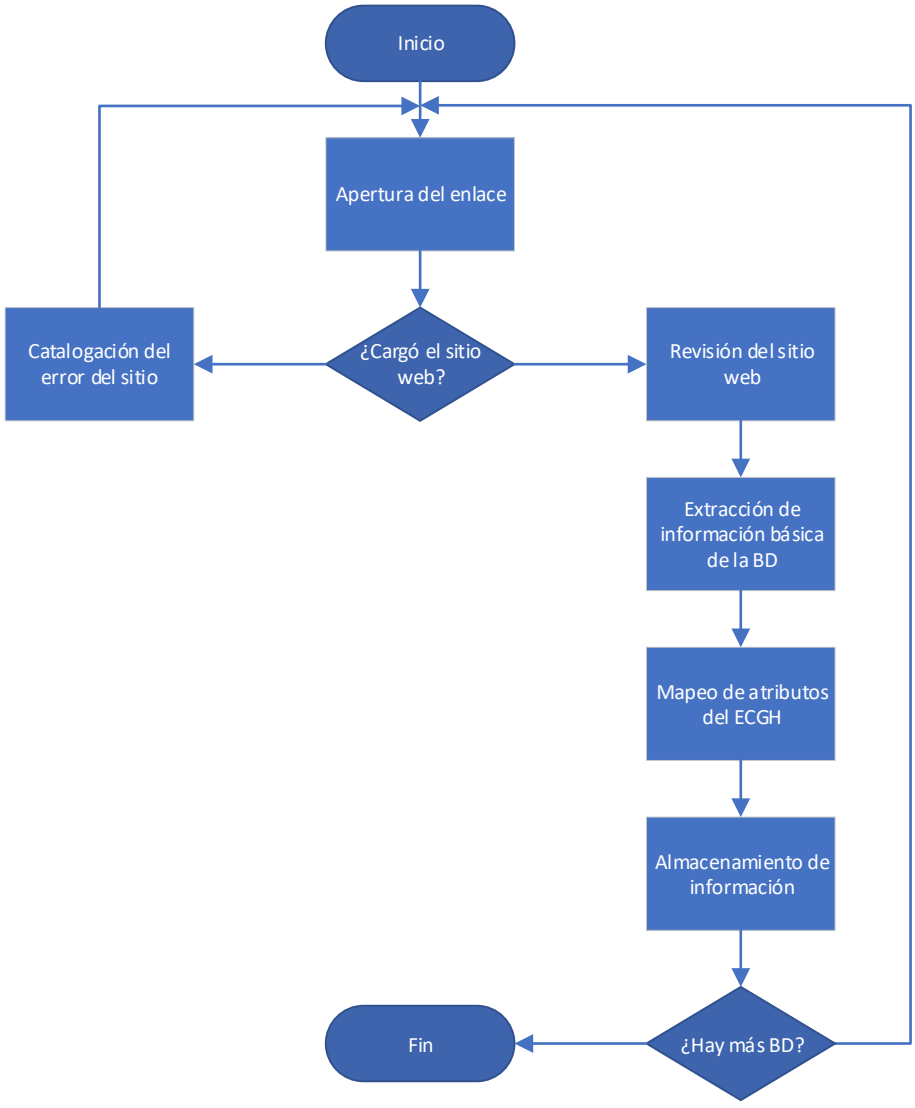
Fuente: Propia

5. Finalmente se consideró los criterios establecidos en la sección 4.1. en donde se filtró cada una de las bases de datos que presentaban información para la especie humana y presenta información relacionada al modelo conceptual quedando así un total de 761 bases de datos genómicas. Es importante mencionar que en este punto las bases de datos que presentaron conectividad y funcionamiento del sitio web se fueron revisando una a una los sitios web de las bases de datos genómicas. Esto se debió a que, en la descripción, nombre o url no se evidenció términos clave que indique información válida para clasificar al sitio.

4.3 TRAZABILIDAD ENTRE MODELO CONCEPTUAL Y BD

Para realizar el mapeo entre el ECGH y cada una de las bases de datos se ha seguido el flujo indicado en la Ilustración 12. Como primer paso para la realizar el mapeo, se tomó el URL de una base de datos genómica y se procedió a buscar los diferentes atributos que se presentaron en el ECGH en la sección 2.

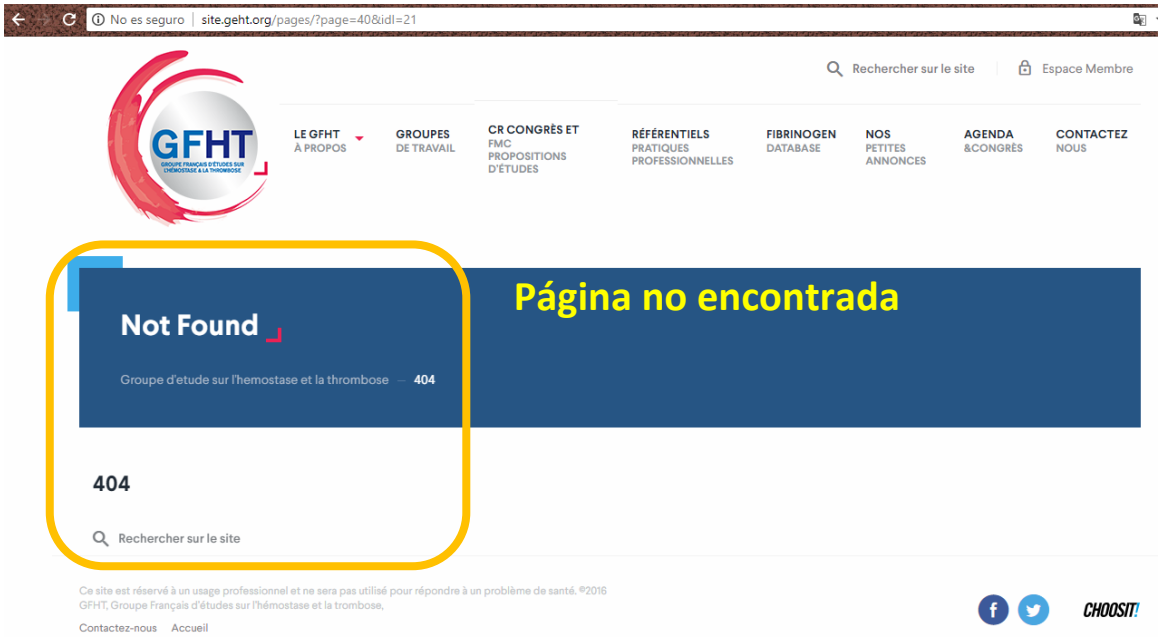
Ilustración 12 Flujo para el mapeo de las BDs y ECGH



Fuente: Propia

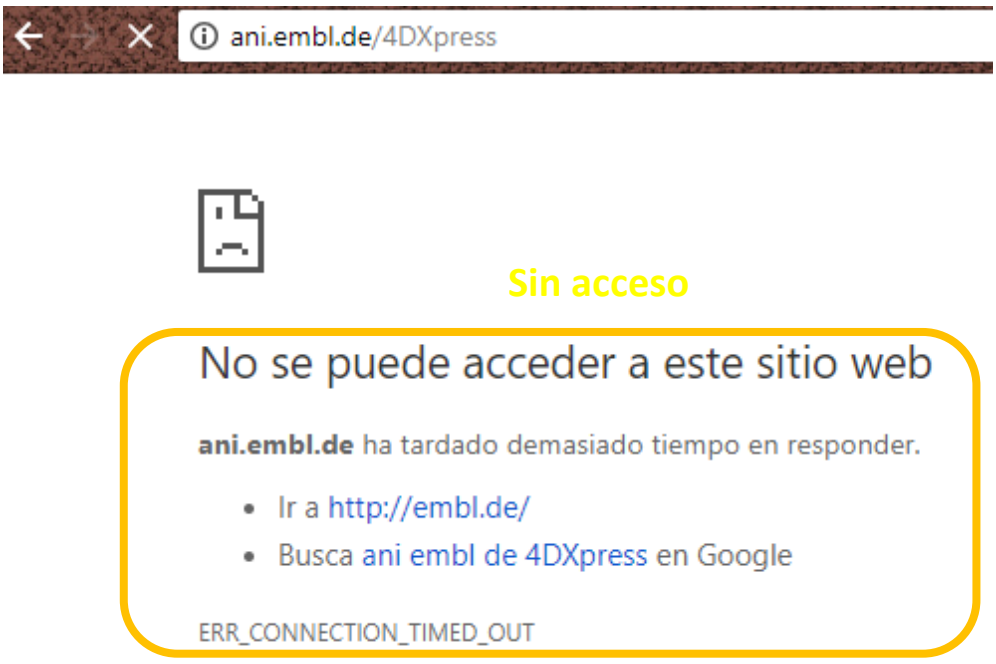
En la Ilustración 13 e Ilustración 14 se puede observar que el sitio no está disponible, por lo que se cataloga como “No encontrado” para el primer caso y “Sin acceso” en el segundo. Algunos sitios web también fueron catalogados como “No permitidos”, “Descontinuados”, “En mantenimiento”, “Servidor no disponible”, “Error de servidor”.

Ilustración 13 Base de datos no disponible



Fuente: Propia

Ilustración 14 Base de datos sin acceso



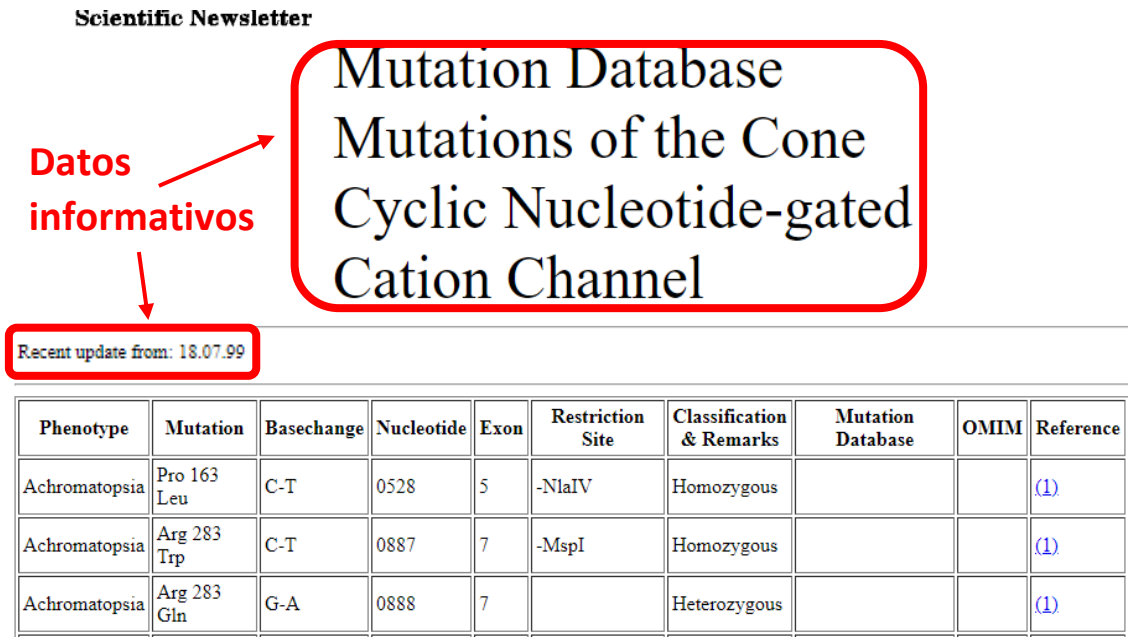
Fuente: Propia

Para el caso de que un sitio web sea accedido correctamente se ha extraído la información básica que no se disponía, como puede ser: "Author/Curators", "Description", "Last update", "Download type". Esta información permite que la persona que esté interesada en detalles de la base de datos pueda de primera mano obtenerla fácilmente. Para este caso se indicará en 2 ejemplos.

EJEMPLO 1: La base de datos genómica contiene información referente a Mutaciones, cuyo nombre es "mutation of the cone cyclic nucleotide-gated cation chanel". Como primer paso se

identifica los elementos principales del sitio, los cuales se pueden visualizar en la Ilustración 16 enmarcados en color rojo.

Ilustración 15 Extracción de datos informativos

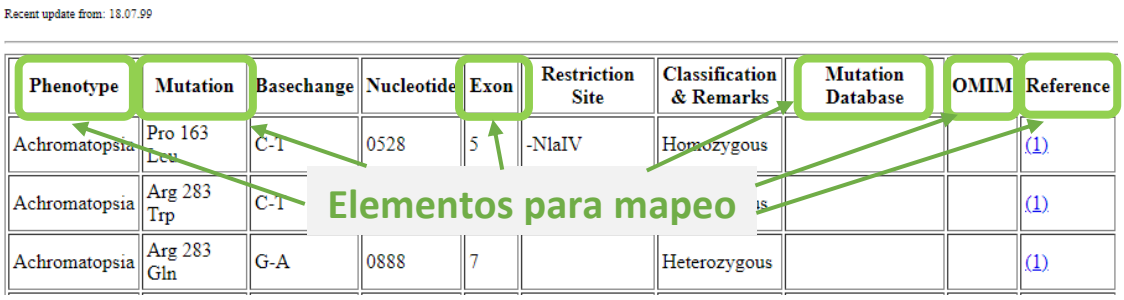


Phenotype	Mutation	Basechange	Nucleotide	Exon	Restriction Site	Classification & Remarks	Mutation Database	OMIM	Reference
Achromatopsia	Pro 163 Leu	C-T	0528	5	-NlaIV	Homozygous			(1)
Achromatopsia	Arg 283 Trp	C-T	0887	7	-MspI	Homozygous			(1)
Achromatopsia	Arg 283 Gln	G-A	0888	7		Heterozygous			(1)

Fuente: Propia

Posteriormente se realiza la identificación de cada uno de los elementos que presenta la base de datos genómica (Ver Ilustración 16), en contraste con el ECGH. Esta información es de gran utilidad ya que en algunos casos se debe considerar que al ser representado un texto por detrás este contiene un enlace que vincula a otros sitios web por lo que se extrae también dicha información.

Ilustración 16 Identificación de elementos para mapeo

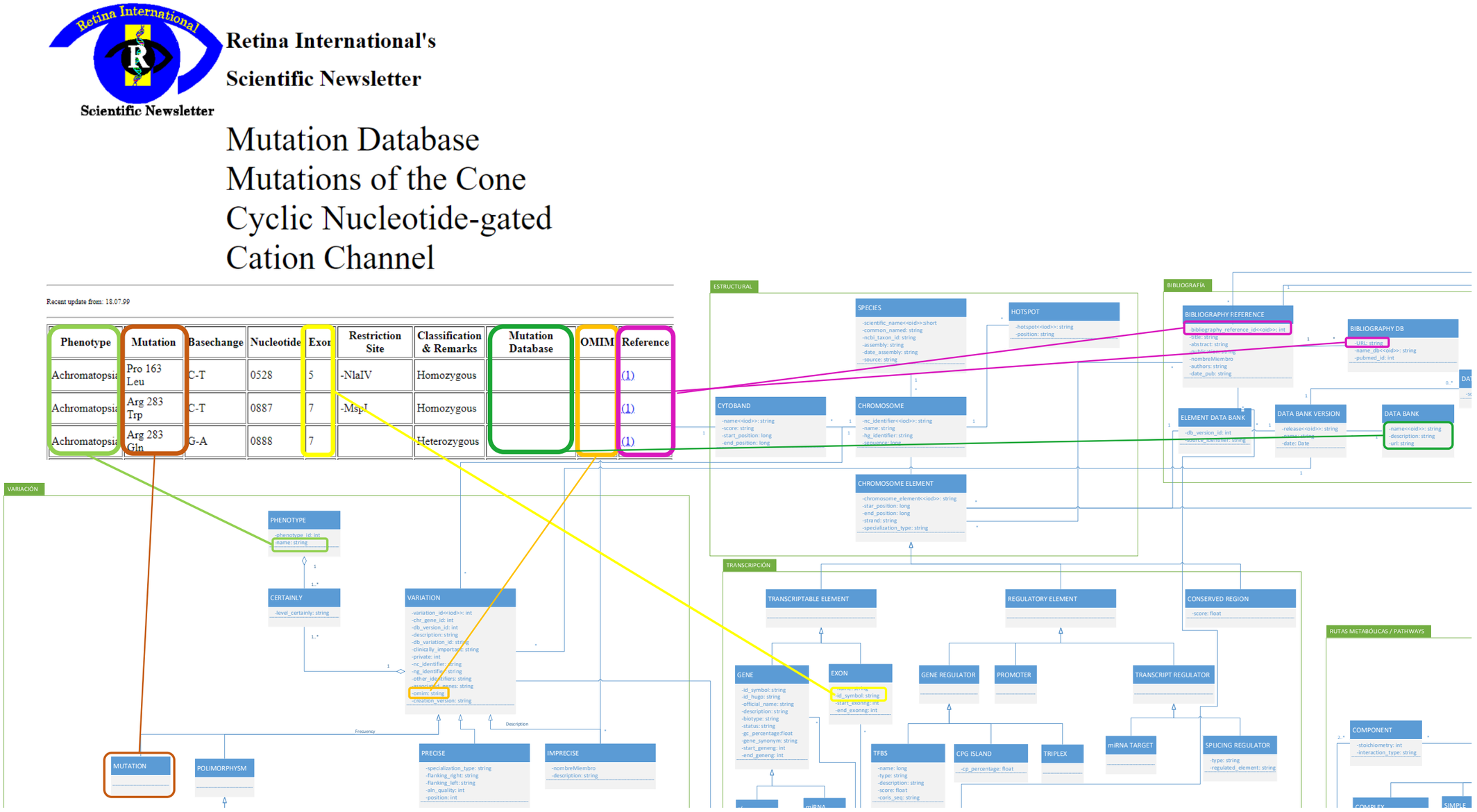


Phenotype	Mutation	Basechange	Nucleotide	Exon	Restriction Site	Classification & Remarks	Mutation Database	OMIM	Reference
Achromatopsia	Pro 163 Leu	C-T	0528	5	-NlaIV	Homozygous			(1)
Achromatopsia	Arg 283 Trp	C-T							(1)
Achromatopsia	Arg 283 Gln	G-A	0888	7		Heterozygous			(1)

Fuente: Propia

Finalmente se realiza el mapeo de manera manual entre cada uno de los atributos que se presentan en el ECGH y los datos presentados en la base de datos (Ver Ilustración 17). Cabe indicar que en el momento de encontrar un elemento que pertenece al ECGH y una BD genómica, la relación que se encuentra se expresa como una especificación estricta, es decir que dicho elemento puede nombrarse como el planteado en el ECGH cuando se refiera a el. Por ejemplo (considerando la Ilustración 17): si se refiere a la información fuente sobre las bases de datos de mutaciones se indicará que estas bases de datos son consideradas como Data Bank, ya que con ese nombre se conoce en el ECG.

Ilustración 17 Mapeo de información de BD y ECGH



Fuente: Propia

Dicha información se va almacenando manualmente en un archivo CSV (Ver Ilustración 18). En la Tabla 7 se puede apreciar el mapeo que se realiza en la Ilustración 17, donde la columna de la izquierda representa a los elementos del ECGH y en la derecha los elementos que se encontraron en la base de datos.

Ilustración 18 Archivo CSV de registro del mapeo

Name DB	URL	AUTHORS	DESCRIPTION	FUENTE	YEAR
SuperSite	http://bioinf-services.charite.de/supersite/		Look at protein structure from a ligand and binding site perspective.	HSLS	
SuperSweet	http://bioinf-applied.charite.de/sweet/	Jessica Ahm	Find information about natural and artificial sweeteners.	HSLS	
SuperTarget	http://insilico.charite.de/supertarget	Hecker, Niko	Find information about drugs.	HSLS	
SuperToxic	http://bioinf-services.charite.de/supertoxic/		A collection of toxic compounds from literature and web sources.	HSLS	
SUPFAM	http://supfam.mbu.iisc.ernet.in/	Mudgal, R.,	This database elucidates the remote relationships found between Pfam families and structural families (SCOP) and uses the evolutionary information inherent of SCOP classification to identify related Pfam familiesRemote relationships between Pfam families.	Oxford, HSLS	
SureChEMBL	https://www.surechembl.org/search/	Hersey, Ann	Chemical compounds extracted from patent documents	Oxford	
SV40 large tumor antigen (T antigen)	http://supernova.bio.pitt.edu/pipaslab/		Search for viruses and plasmids expressing mutant forms of the Simian virus 40 (SV40) large T antigen.	HSLS	
SVMHC	http://www-bs.informatik.uni-tuebingen.de/Services/SVMHC/		Predict MHC-binding peptides.	HSLS	
SVM-Prot	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi		Classify a protein into functional family from its primary sequence.	HSLS	

Fuente: Propia

Tabla 7 Mapeo de elementos del ECGH y BD genómica

MUTATION OF THE CONE CYCLIC NUCLEOTIDE-GATED CATION CHANEL	**ELEMENTOS DEL ESQUEMA CONCEPTUAL DEL GENOMA HUMANO**	
	CLASE	***ATRIBUTO***
PHENOTYPE	Phenotype	name
MUTATION	Mutation	
EXON	Exon	Id_symbol
MUTATION DATABASE	Data Bank	name description url
OMIM	Variation	omim
REFERENCE	Bibliography reference	bibliography_regerence_id
REFERENCE	Bibliography DB	URL

Fuente: Propia

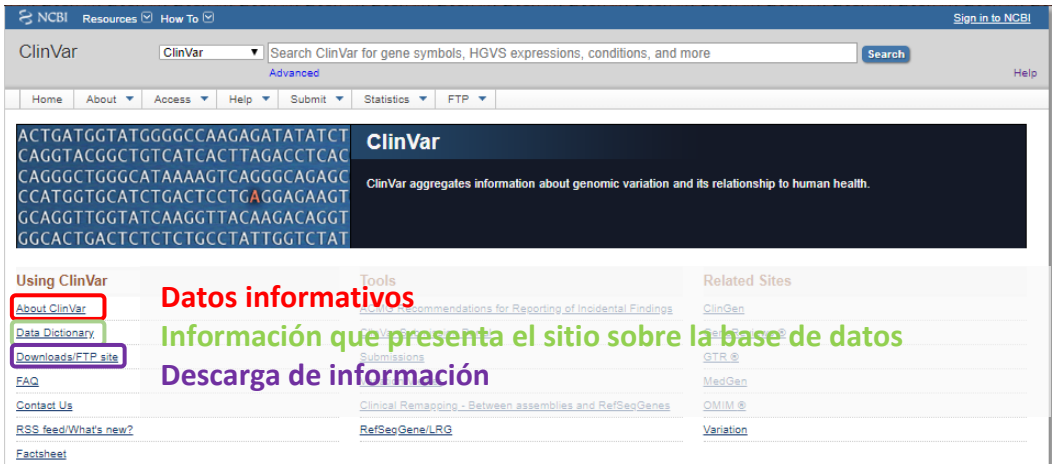
EJEMPLO 2: La base de datos genómica a explicarse es la de Clinvar, la cual pertenece a NCBI. Este sitio es diferente al del ejemplo 1 ya que no da un listado inicial de la información, sino que en base a la búsqueda de términos como:

- símbolos de genes, por ejemplo: PTEN
- expresiones HGVS, por ejemplo: NM_000314.4:c.395G>T
- cambios proteínicos, por ejemplo: G132V
- número rs, por ejemplo: rs180177042
- nombre de enfermedades, por ejemplo: PTEN hamartoma tumor syndrome

- localización de cromosomas, por ejemplo: 10[chr] AND 89623000:89730000[chrpos37] en la cual se estaría buscando las variantes en el cromosoma 10 entre las posiciones 89623000 y 89730000 según GRCh37 (chrpos37)

permite el despliegue de la información que se encuentra en la base de datos. En la Ilustración 19 se puede observar que en la portada del sitio se puede navegar para encontrar información. En el color rojo información como descripción del sitio web, autores, actualización de las bases de datos. En el color violeta información sobre la descarga de información de la base de datos y posible manejo de APIs. En el color verde información sobre la estructura del sitio lo cual puede ayudar y simplificar de cierta manera el mapeo revisando cada uno de los datos que muestra en los reportes.

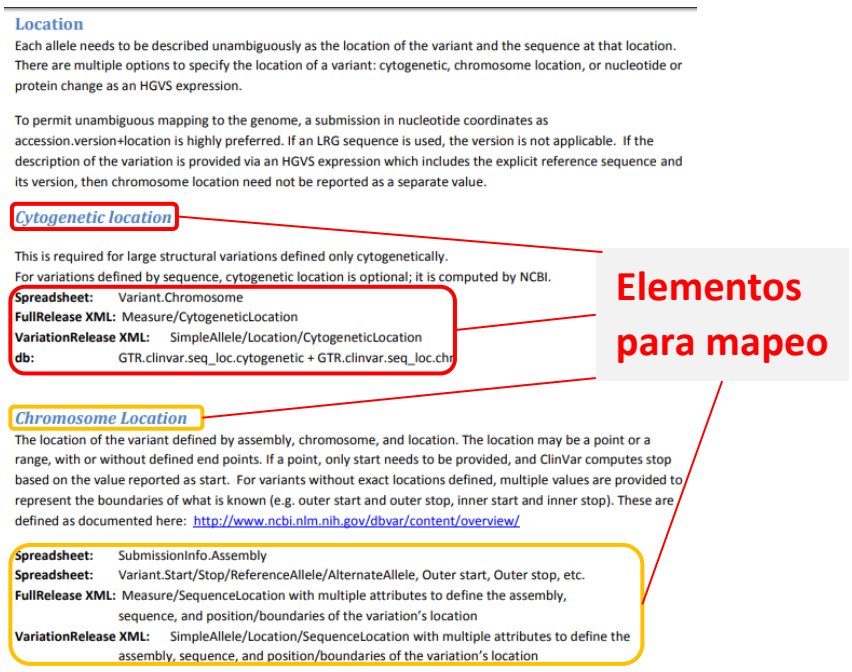
Ilustración 19 Extracción básica de información



Fuente: Propia

En lo que se refiere al mapeo de información, en la Ilustración 20 se puede visualizar que mediante el diccionario de datos presentado se realiza la búsqueda de cada uno de los atributos que forman parte del ECGH.

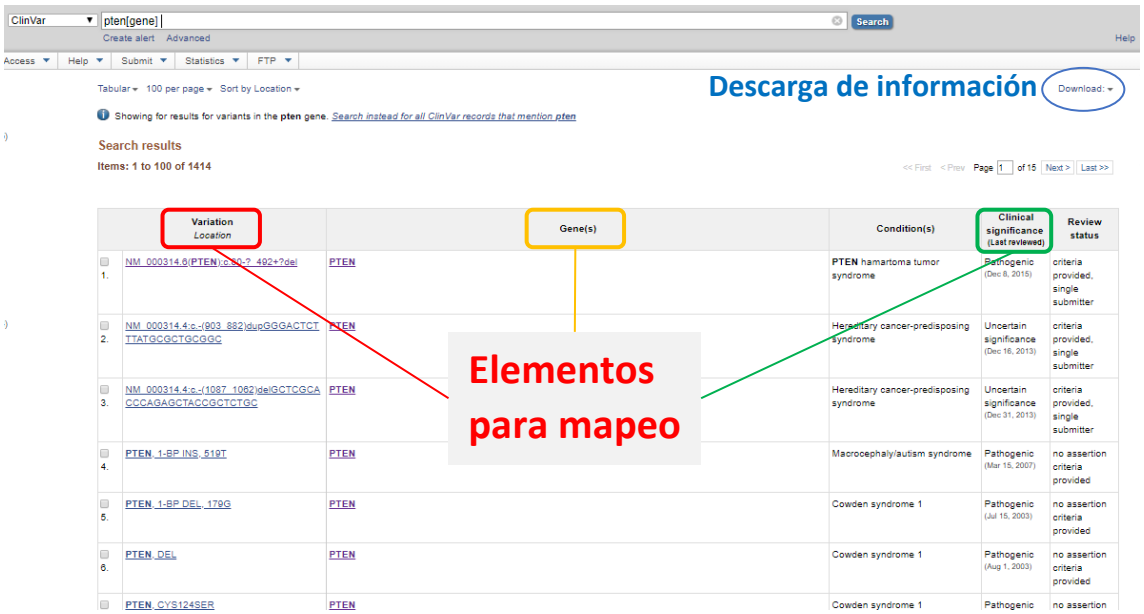
Ilustración 20 Búsqueda de atributos en diccionario de datos



Fuente: Propia

Adicionalmente esto se puede verificar accediendo a la búsqueda de información como se indicó inicialmente por diferentes entradas, en este caso se lo realizará por la búsqueda de un gen (Ver Ilustración 21), aquí se puede apreciar que en los resultados se refleja información que deseamos mapear la cual se ha enmarcado.

Ilustración 21 Búsqueda de información por gen



Fuente: Propia

De manera similar en la Ilustración 22, cuando se realiza la búsqueda de información en base al nombre de la enfermedad, en este caso hemos hecho según "*PTEN hamartoma tumor syndrome*". Se puede apreciar en lo enmarcado que existe información útil para el mapeo correspondiente con el ECGH.

Ilustración 22 Búsqueda por nombre de enfermedad

Fuente: Propia

Una vez que se ha identificado la información base se explora tanto las variaciones como los genes (ya que llevan a otras páginas). En la Ilustración 23 se indica el mapeo que se va realizando desde las diferentes páginas que se han encontrado en la navegación con respecto al ECGH. Se ha señalado con diferentes colores a cada uno de los elementos para que sea más fácil la ubicación de estos. Para una fácil comprensión en la Tabla 8 se indican las relaciones que se han encontrado en base a la Ilustración 23, aquí se puede apreciar que para las diferentes clases y

atributos existen elementos que la base de datos presenta no obstante en cada uno de ellos pueden que en la base de datos se indiquen con diferentes nombres o de manera diferente al que fue concebido el ECGH.

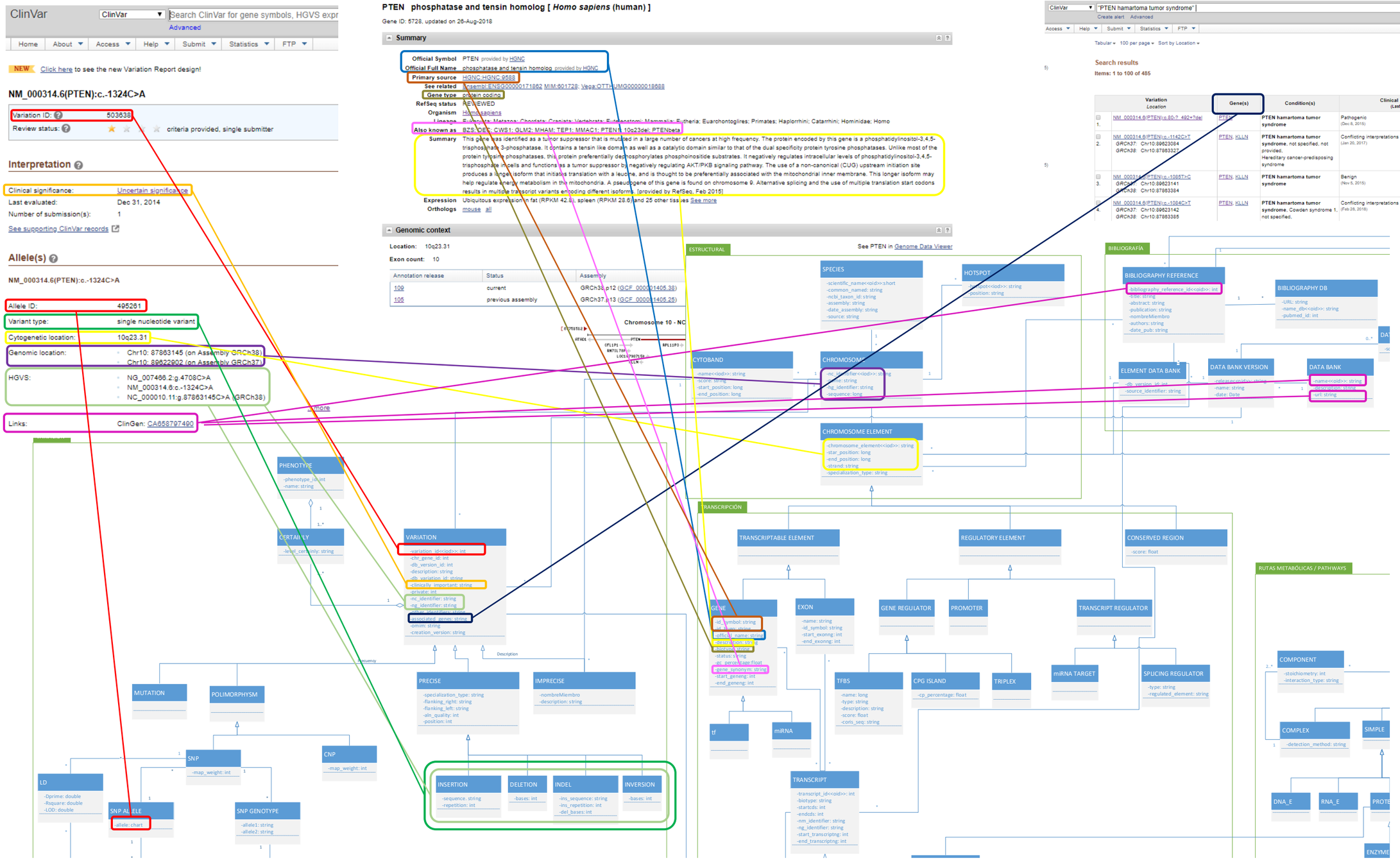
Tabla 8 Mapeo de elementos del ECGH y BD genómica

CLINVAR	ELEMENTOS DEL ESQUEMA CONCEPTUAL DEL GENOMA HUMANO	
	CLASE	***ATRIBUTO***
ALLELE ID	SNP_ALLELE	alelle
VARIATION ID **CLINICAL SIGNIFICANCE** **HGVS** **HGVS** **GENE(S)**	VARIATION	Variation_id Clinically_important Nc_identifier Ng_identifier Associated_genes
VARIANT TYPE **HGVS**	INSERTION	Sequence Repetition
VARIANT TYPE **HGVS**	DELETION	Bases
VARIANT TYPE **HGVS**	INDEL	Ins_ sequence Ins_repetition Del_bases
VARIANT TYPE **HGVS**	INVERSION	bases
CYTOGENETIC LOCATION	CHROMOSOME ELEMENT	Chromosome_element Star_position End_position strand
PRIMARY SOURCE **PRIMARY SOURCE** **OFFICIAL SYMBOL/OFFICIAL FULL NAME** **SUMMARY** **GENE TYPE** **ALSO KNOW AS**	GENE	Id_symbol Id_hugo Official_name Description Biotype Gene_synonym
GENOMIC LOCATION	CHROMOSOME	Nc_identifier Name Hg_identifier sequence
CLINGEN	BIBLIOGRAPHY REFERENCE	Bibliography_reference
CLINGEN	DATA BANK	Name url

Fuente: Propia

Los elementos encontrados van a su vez actualizándose y alimentando el archivo CSV como resultado del mapeo. Es relevante indicar que dependiendo de cada uno de los sitios web en algunos casos la descarga de información puede encontrarse luego de realizar búsquedas o directamente en apartados de descarga de información en donde inclusive se indica si la base

Ilustración 23 Mapeo de información con el ECGH



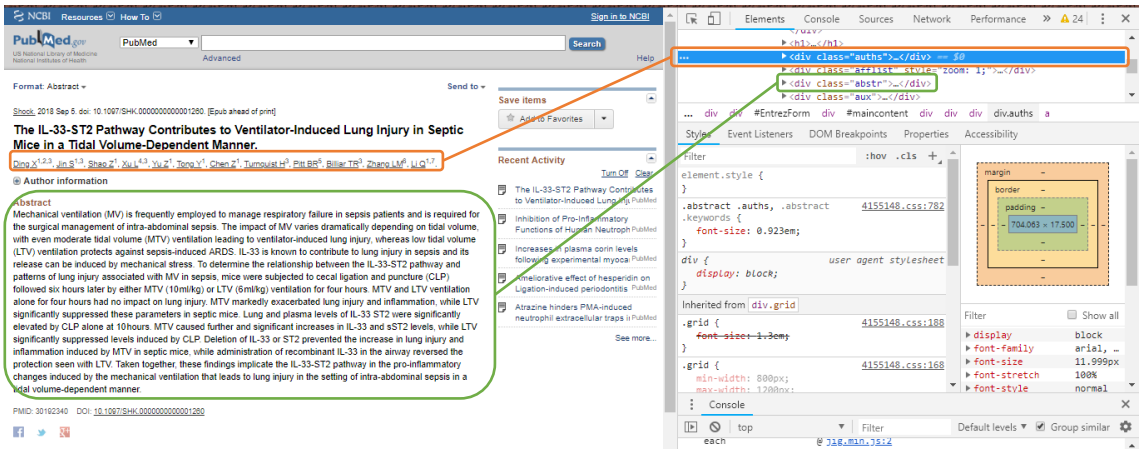
Fuente: Propia

de datos genómica permite o contiene algún servicio mediante el cual se pueda realizar la descarga de datos. Esta información es considerada también almacenado en el csv.

Posteriormente con la información recolectada de las diferentes bases de datos genómicas habilitadas se puede obtener una trazabilidad. La persona que desee encontrar las bases de datos que presentan información sobre la vista estructurada por ejemplo podrán tener un listado de todas ellas, así mismo si es el caso de querer saber qué información presenta una base de datos, pues basta con seleccionar la base de datos genómica y como resultado se le presentará los atributos que pudieron mapearse en base al ECGH.

Adicionalmente para enlazar encontrar la sección en los casos de que se presente solamente a manera de texto la información que se busca se ha considerado la estructura html que presenta el sitio web de la base de datos genómica. En la Ilustración 24 se presenta un ejemplo en el cual dentro de la base de datos de PubMed se encuentra los autores del artículo científico para lo cual en la estructura html del sitio se puede apreciar que esta información está dentro de la sección “auths”. De la misma manera para el caso del abstract se ha encontrado la sección en la que se encuentra dentro del html, siendo asi que se llama la sección “abstr”. Por lo que toda esta información se agregará al archivo csv como dato referencial del atributo encontrado.

Ilustración 24 Trazabilidad mediante elementos de página web



Fuente: Propia

4.4 PRESENTACIÓN DE RESULTADOS

En esta sección se presentan las herramientas utilizadas para la visualización de datos, así como también la aplicación resultado de la exploración de BD genómicas. Además, se ha agregado una sección de estadísticas con los resultados de la exploración.

4.4.1 TECNOLOGÍAS EMPLEADAS

Para la realización de la página web se utilizó:

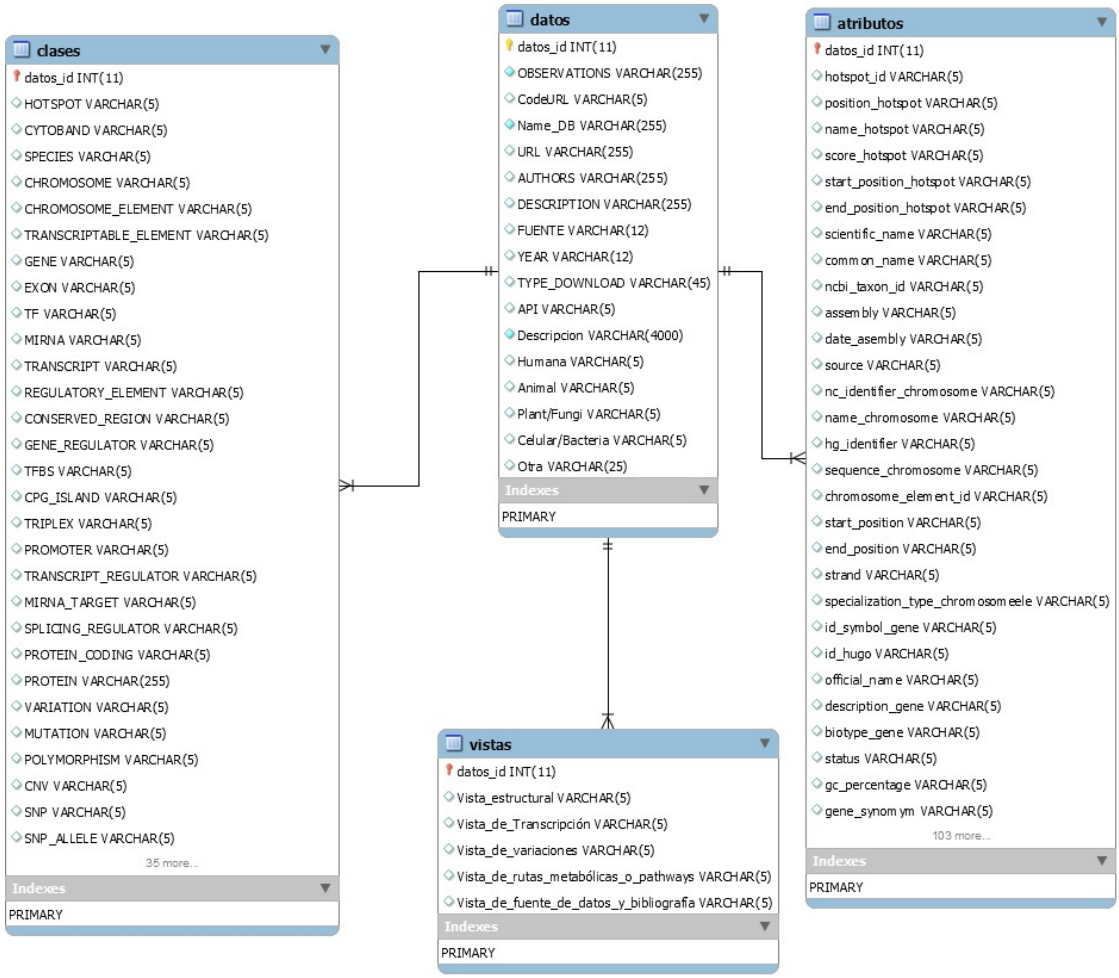
- Framework Django versión 2.1.1, para la realización del visualizador de resultados de bases de datos genómicas.

- Python 3.6.5, como herramienta auxiliar para la programación dentro de Django.
- Se migró el archivo csv a una base de datos en MySQL versión 5.7 como repositorio de la información obtenida.

4.4.2 SITIO WEB

Una vez que se ha completado el archivo csv como información fuente del mapeo entre las bases de datos genómicas y el ECGH, se lo ha ingresado a una base de datos en MySQL para facilitar la presentación de información y obtener una trazabilidad en las búsquedas. En la Ilustración 25 se puede apreciar el diagrama de la base de datos que se ha generado importando la información del csv. El cual consta de las vistas genómicas, clases del ECGH, atributos de las diferentes clases e información de las bases de datos genómicas.

Ilustración 25 Diagrama de base de datos

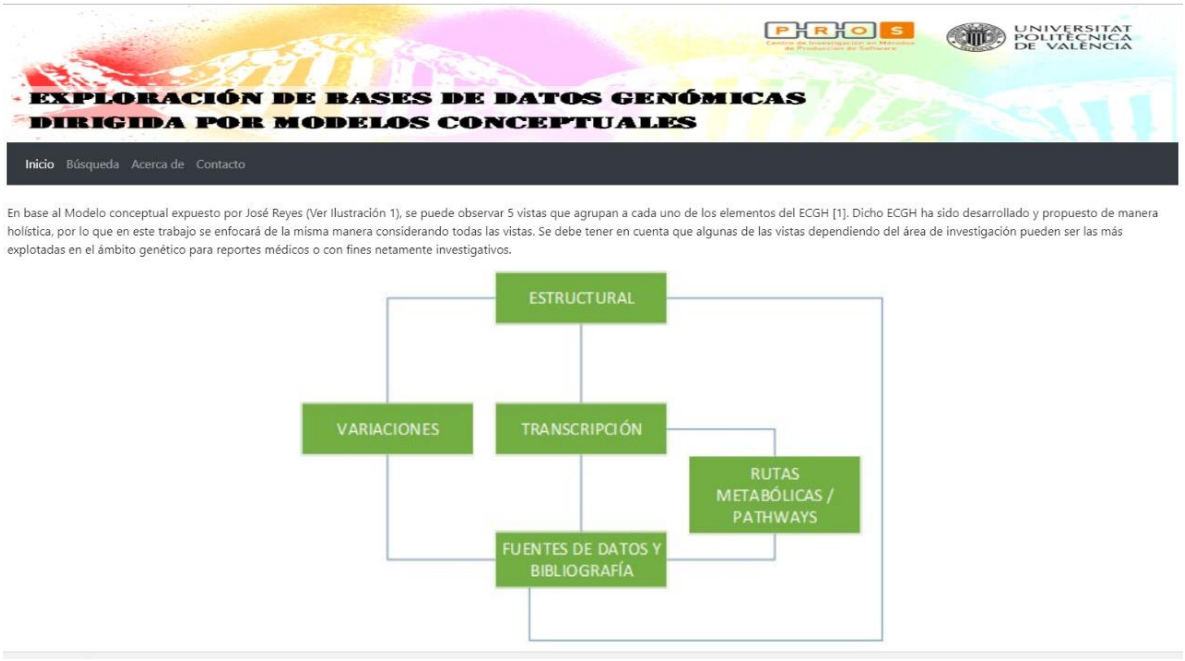


Fuente: Propia

Para una mejor visualización de la información se creó un sitio web en Django que permite realizar las consultas referentes a que bases de datos presentan información en base a las vistas, clases y

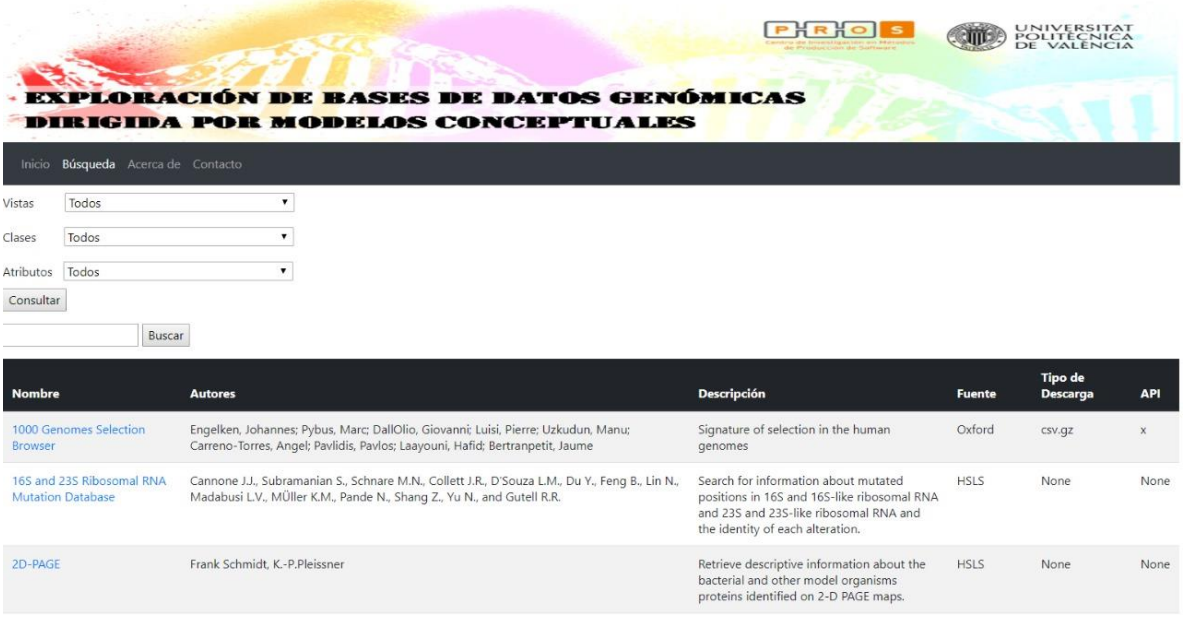
atributos respectivamente (Ver Ilustración 26). En la Ilustración 27 se puede observar que el sitio creado permite una fácil localización de los elementos a buscarse.

Ilustración 26 Herramienta para la visualización de resultados



Fuente: Propia

Ilustración 27 Visualizador creado para búsquedas



Fuente: Propia

Para el caso de requerirse conocer las bases de datos genómicas que presentan información sobre alguna vista (Ver Ilustración 28) se selecciona la vista deseada y se da clic en buscar, presentándose

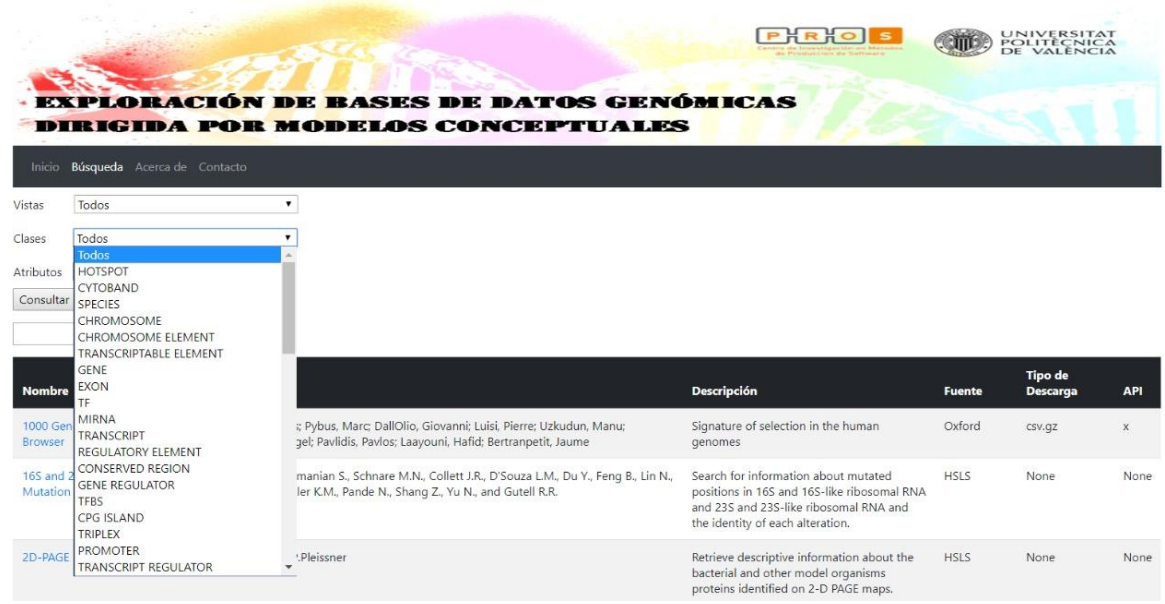
un listado de bases de datos que poseen esta información. El listado de bases de datos permite acceder a los sitios web y a su vez se indica los formatos de descarga de información, así como si permite la utilización de APIs para el manejo de la misma.

Ilustración 28 Visualizar BD genómicas por Vistas

Fuente: Propia

Si la búsqueda va enfocada a conocer sobre las bases de datos que están asociadas con alguna clase (Ver Ilustración 29) se selecciona la clase y se da clic en buscar. Como resultado se presenta un listado de las bases de datos genómicas que indican información sobre dicha clase, permitiendo el acceso a la DB, así como también indicando los formatos de descarga de datos y utilización de API.

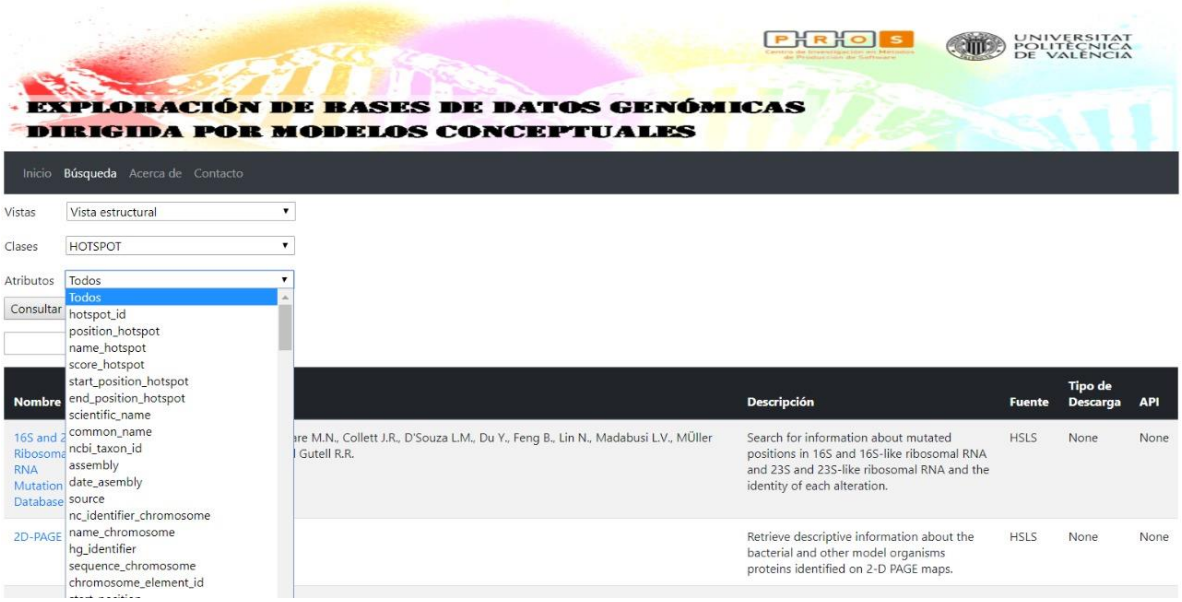
Ilustración 29 Visualizar BD genómicas por Clases



Fuente: Propia

En caso de que la búsqueda se realice por algún atributo (Ver Ilustración 30), se selecciona el atributo y se presentan los resultados de manera similar a las otras búsquedas.

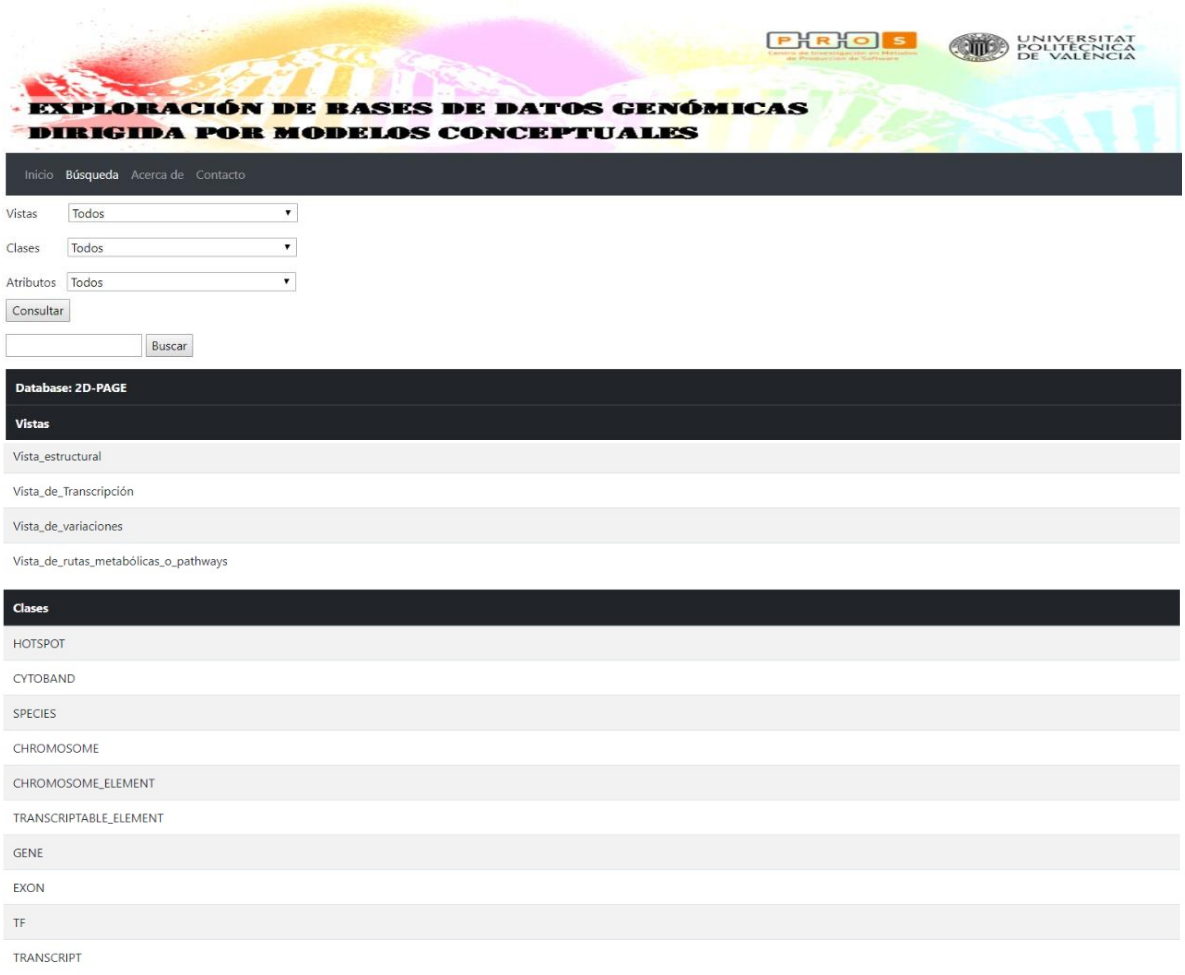
Ilustración 30 Visualizar BD genómicas por Atributos



Fuente: Propia

Para el caso inverso para lograr la trazabilidad se pueden realizar búsquedas considerando una BD genómica (Ver Ilustración 31), como resultado presentará información básica de la base de datos así como las vistas, clases y atributos que están asociados según el ECGH.

Ilustración 31 Visualizar contenido de BD genómica

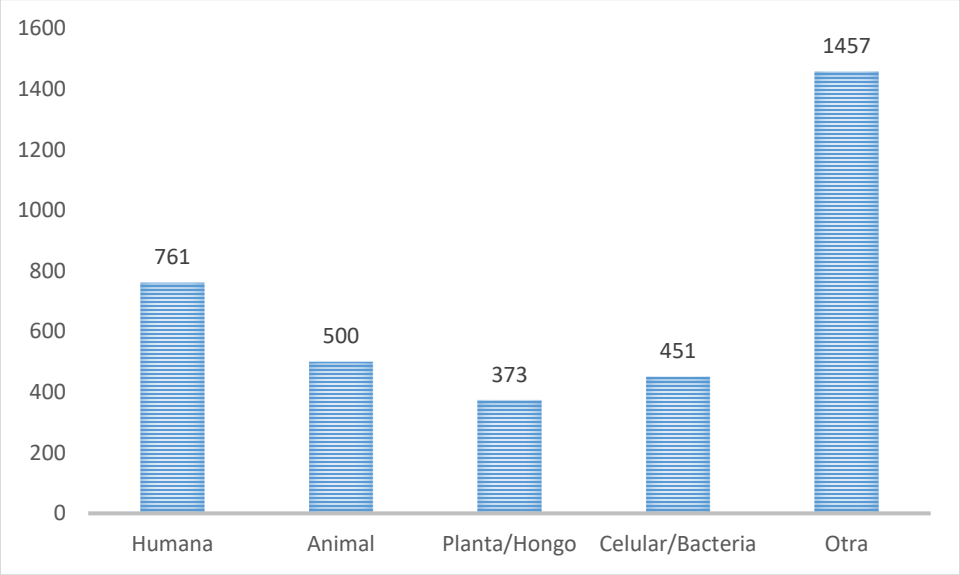


Fuente: Propia

4.4.3 DATOS ESTADÍSTICOS

La exploración de las diferentes bases de datos genómicas(Ver Ilustración 32) ha permitido evaluar que se encontraron 3542, de las cuales el 21.49% corresponde a la especie humana, el 14.12% a animales, el 10.53% a plantas u hongos, el 12.73% a células o bacterias y finalmente el 41.13% está disperso en diferentes subcategorías. En la Tabla 9 se puede apreciar la subcategorización que se obtuvo dentro de la subclasificación para especies. Se menciona como punto importante que este listado también forma parte de la base de datos de consulta.

Ilustración 32 BD genómicas por especie



Fuente: Propia

Tabla 9 Otras especies

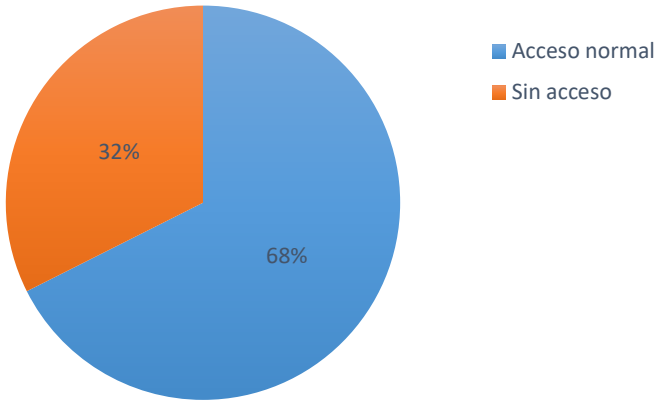
OTRAS CLASIFICACIONES	NO.	%
ALLERGEN	1	0,11
AMINOACIDE	3	0,32
ANTIBODIES	1	0,11
BIOLOGICAL ACTIVITY	1	0,11
BIOMARKER	1	0,11
CALCULUS	2	0,21
CARBOHYDRATE	1	0,11
CDNA	3	0,32
CLINICAL CRITERIA	2	0,21
CLÚSTER	17	1,81
COMPARATIVE DATA	112	11,95
CONVERSOR	4	0,43
DATABASE	1	0,11
DNA REPLICATION	1	0,11
DRAWING EULER DIAGRAM	1	0,11
DRUG	33	3,52
ENZYME	7	0,75
FORO	1	0,11
GENE QUANTIFICATION	1	0,11
GENECHIPS	1	0,11
GRAPHIC	22	2,35
HAPLOTYPE	1	0,11
HERIETABILITY	1	0,11
MICROARRAY	25	2,67

OTRAS CLASIFICACIONES	NO.	%
MICROBIAL	30	3,20
MOLECULE	96	10,25
ONTOLOGY	18	1,92
PATENT SEQUENCE	1	0,11
PATHOGEN	3	0,32
PEPTIDE	21	2,24
PHYLOGENETIC TREES	1	0,11
PICK PRIMER	12	1,28
PLACES	1	0,11
PREDICTION	143	15,26
PROTOCOLS	5	0,53
REACTIONS	1	0,11
RNA	29	3,09
SIMULATION	3	0,32
SOFTWARE	96	10,25
STATISTIC	15	1,60
STRUCTURE	139	14,83
SWEETENERS	1	0,11
TOOL	31	3,31
TOXIN	1	0,11
TRADUCTOR	1	0,11
VACCINE	2	0,21
VECTOR	1	0,11
VIRUS	25	2,67
WIKI	18	1,92
TOTAL	937	100,00

Fuente: Propia

Con el mapeo de cada una de las bases de datos ha permitido verificar que las URL proporcionadas inicialmente lleguen a ser efectivamente las correctas, ya que en el momento de revisar el contenido de estas se encontró que muchas de ellas no eran accesibles o su enlace estaba roto. Como se puede ver en la Ilustración 33 que corresponde al 100% de la exploración de bases de datos genómicas, el 32% representa a 1148 bases de datos genómicas que no han sido accedidas.

Ilustración 33 Acceso a BD genómicas



Fuente: Propia

Las posibles causas al acceso fallido son las presentadas en la Tabla 10. Así mismo el 68% Ilustración 33) de bases de datos si pudieron accederse o encontrar un URL válido y es de las que se ha podido realizar el mapeo correspondiente.

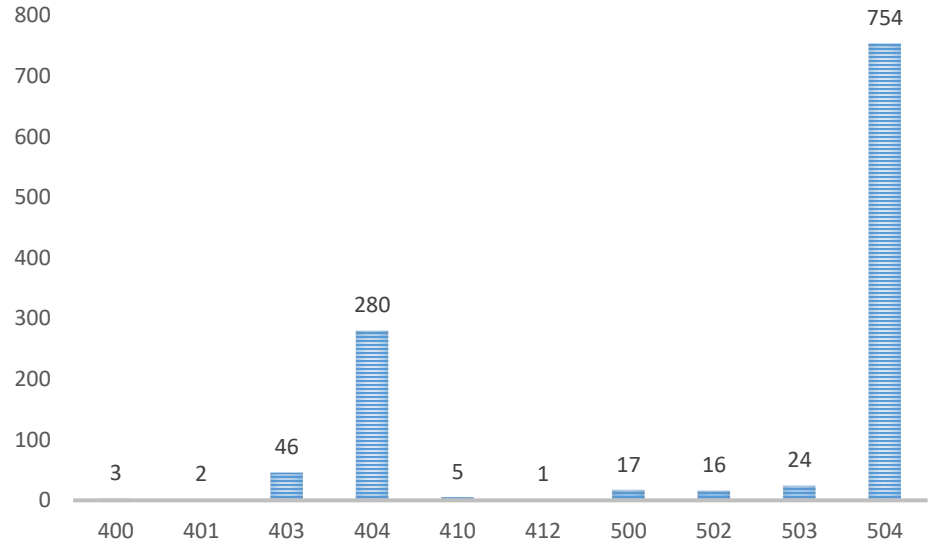
Tabla 10 Errores en carga de sitios web de DB genómicas

DESCRIPCIÓN	CÓDIGO	NO.	%
PETICIÓN DEL SITIO CON ERROR	400	3	0,26
SIN AUTORIZACIÓN	401	2	0,17
USUARIO SIN PERMISOS	403	46	4,01
RECURSO NO DISPONIBLE	404	280	24,39
EL RECURSO SE HA QUITADO	410	5	0,44
PETICIÓN NO CUMPLIDA	412	1	0,09
SITIO CAÍDO	500	17	1,48
GATEWAY NO DISPONIBLE	502	16	1,39
SERVIDOR CONGESTIONADO O EN MANTENIMIENTO	503	24	2,09
TIME OUT EN EL SERVIDOR	504	754	65,68
	Total	1148	100,00

Fuente: Propia

En la Ilustración 34 se pueden observar las causas de la no conexión con las bases de datos genómicas, en donde el error por exceso en la espera de respuesta del servidor (754 casos) corresponde al 65.68 % siendo la mayor causa, seguido por el 24.39% que representa a 280 casos. Una de las soluciones en el momento de verificación de acceso a los sitios fue el de ingresar a la dirección base de la URL en donde se buscó el proyecto o base de datos indicado.

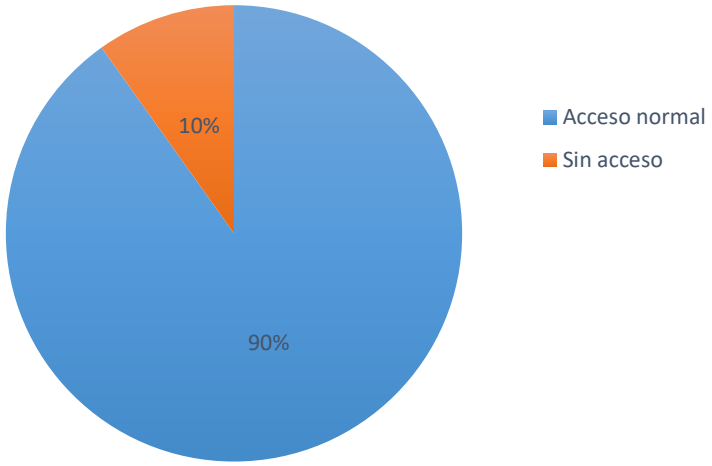
Ilustración 34 Causas de no conexión



Fuente: Propia

En la revisión de las diferentes BD genómicas se pudieron clasificar en base a la información que maneja con respecto al tipo de especie, para lo cual se observó en la Tabla 9 que para la especie humana (Homo Sapiens-Human) se tiene 761 BD, de las cuales se puede apreciar en la Ilustración 35 que el 10% no pudieron ser accedidas por diferentes causas.

Ilustración 35 Acceso a las BD genómicas: Especie Humana

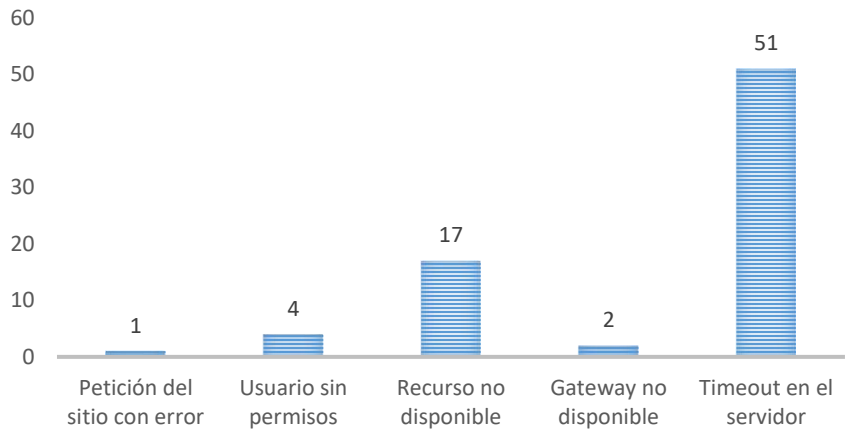


Fuente: Propia

En la Ilustración 36 se pueden observar las causas del no acceso a las 75 bases de datos genómicas correspondientes a la especie humana, donde hay 51 casos que corresponden al 68% de exceso en tiempos de espera de respuesta del servidos, 17 casos corresponden al 22.67% de recursos que se han quitado, seguido por 4 casos (5.33%) cuando el usuario no tiene permisos, 2 (2.67%) casos

donde el Gateway no está disponible y finalmente 1 (1.33%) caso cuando hubo una petición al sitio con algún error.

Ilustración 36 Causas de no acceso a las BD genómicas: Especie Humana



Fuente: Propia

No obstante es indispensable indicar que en muchos sitios explorados se encontraron páginas que en su contenido no correspondían al indicado, o se encontraba en mantenimiento, esto se puede apreciar en la Tabla 10, en la cual se presenta la catalogación que se ha dado a los sitios que han presentado ciertas irregularidades, posterior a la carga normal de la página web de la base de datos genómica.

Tabla 11 Errores encontrados en sitios con conexión

IRREGULARIDADES	NO.	%
AUTHORIZATION REQUIRED	6	16,67
DATABASE ERROR	2	5,56
DISCONTINUED	1	2,78
INTERNAL ERROR	6	16,67
INTERNAL SERVER ERROR	3	8,33
MAINTENANCE	7	19,44
REDIRECTION	1	2,78
RETIRED	1	2,78
UNAVAILABLE	8	22,22
WEBSITE CLOSED	1	2,78
TOTAL	36	100,00

Fuente: Propia

De las 725 bases de datos genómicas que finalmente pudieron encontrarse habilitadas tanto en acceso como en información, se pudo apreciar que algunas de ellas permiten en sus sitios web la

descarga de archivos o recursos en diferentes formatos los cuales se detallan en la Tabla 12, teniendo 240 DB genómicas que permiten la descarga. De manera similar algunos sitios web han incorporado herramientas que contienen un conjunto de funciones y procedimientos, las cuales permiten la creación de aplicaciones para el acceso a los datos que contiene. Estas herramientas han sido catalogadas como APIs, la Tabla 13 muestra que el 9.33% permite el uso de ellas, mientras que el 90.67% no tiene incorporado estas herramientas o servicios.

Tabla 12 DB genómicas que permiten descarga de recursos (Especie Humana)

PERMITE DESCARGA DE RECURSOS	NO.	%
SI	240	34,99
NO	446	65,01
TOTAL	686	100,00

Fuente: Propia

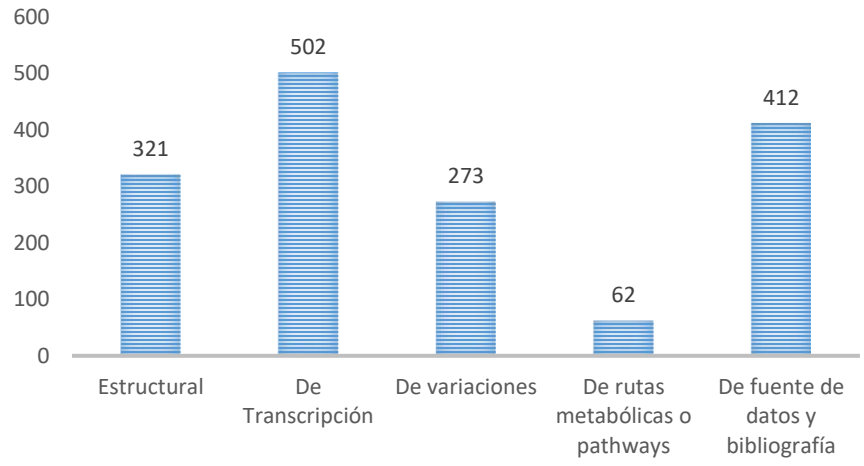
Tabla 13 DB genómicas que incorporan APIs (Especie Humana)

PERMITE USO DE API	NO.	%
SI	71	9,33
NO	690	90,67
TOTAL	761	100,00

Fuente: Propia

Finalmente se ha querido visibilizar el conteo de las bases de datos que se encuentran agrupadas por las 5 vistas genómicas consideradas en el ECGH. La Ilustración 37 indica que la agrupación más grande de DB es en la vista de transcripción, ya que existen 502 BD. Siendo la menor presencia de BD para la vista de rutas metabólicas o pathways con 62 DB genómicas.

Ilustración 37 BD agrupados por vistas



Fuente: Propia

Como complemento de este trabajo, se incorporó para las consultas las bases de datos que no cumplieron con los criterios de selección, de las cuales puede obtenerse nombre, descripción, autor y URL (Ver Tabla 14). Por este motivo a pesar de que estos no fueron mapeados con el ECGH, forman parte de un listado adicional en el que se ha categorizado por especie. Se tiene un total de 2781 bases de datos genómicas.

Tabla 14 Otras bases de datos genómicas

ESPECIE	NO.	%
ANIMAL	500	17,98
PLANTA/HONGO	373	13,41
CELULAR/BACTERIA	451	16,22
OTRA	1457	52,39
TOTAL	2781	100,00

Fuente: Propia

5 CONCLUSIONES

En esta sección se presentarán las diferentes conclusiones que se ha llegado al finalizar el trabajo de fin de máster. A partir de cada uno de los objetivos planteados se indicará cuáles han sido los problemas encontrados. Adicionalmente se planteará algunos trabajos futuros que pudiesen surgir a partir de los resultados de este trabajo.

5.1 PROBLEMAS ENCONTRADOS

En la elaboración del trabajo de fin de máster se pudieron detectar diferentes problemas, los cuales se detalla a continuación:

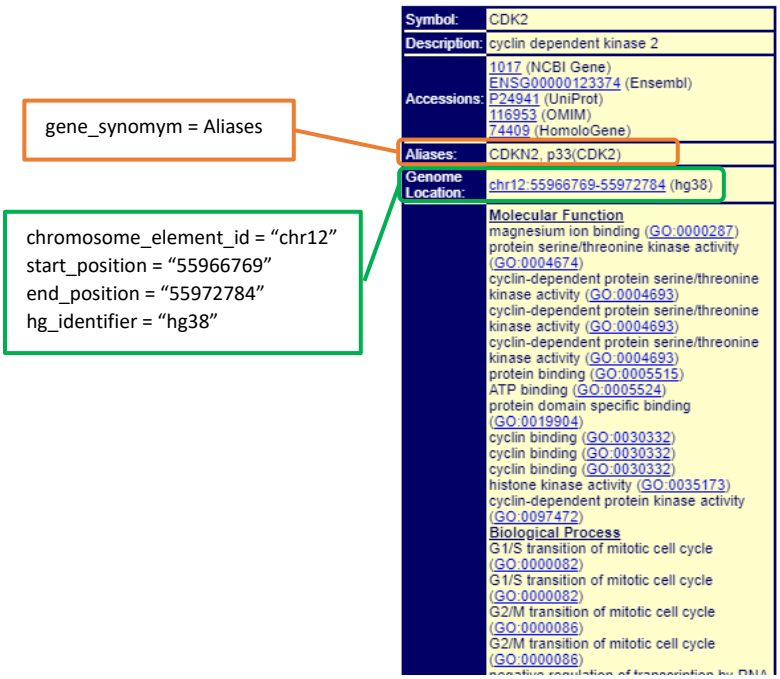
- Uno de los principales problemas encontrados es el llamado “Caos de datos genómicos”. Esto se lo considera muy relevante ya que la información presentada en las diferentes fuentes mantiene un grado de dispersión, heterogeneidad, redundancia y muchas de las veces inconsistencia, por lo que es complicado el tratamiento de los datos que presentan. Es así como, el dominio del tema es fundamental, en conjunto con el conocimiento de la estructura de datos que presenta cada uno de los sitios web correspondientes a las bases de datos orientadas al genoma humano.

- Si bien cada repositorio presenta una estructura diferente, también es el caso de las nomenclaturas que se disponen y presentan en el momento de consulta de la información, por lo que el investigador o persona que recaba dicha información debe poseer conocimientos bases que lo puedan guiar y conducir en la investigación. Por este motivo, la recopilación y mapeo de cada una de las bases de datos genómicas con el esquema conceptual genómico se realizó de forma manual, no descartando datos importantes por cambio de nomenclaturas o abreviaturas que pudiesen dar a confusión. Como consecuencia de estos “sinónimos” en terminología genética se visibiliza el caos de datos genómico por la heterogeneidad de la información. Es en este punto gracias a la versatilidad que posee el ECGH, este permite el mapeo directo de cada atributo y si fuese el caso agregar algún elemento que pueda ser de utilidad y relevancia para el investigador o médico.

- A continuación, se presentan algunos de los problemas encontrados en la exploración de bases de datos genómicas:

  EJEMPLO 1: En la base de datos genómica BioGPS (http://biogps.org) se ha detectado en la tabla que el campo “Genome location” pertenece en el ECGH a diferentes atributos como se aprecia en la Ilustración 38 enmarcado en color verde, así mismo en color naranja “Aliases” que se asocia con “gene_synomym” dentro del ECGH.

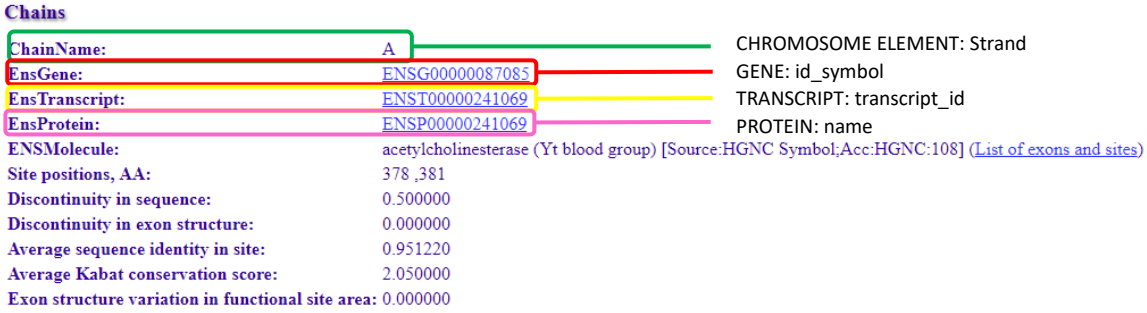
Ilustración 38 Ejemplo problema 1



Fuente: Propia

EJEMPLO 2: En la base de datos genómica “SitEX” (http://www-bionet.sscc.ru/sitex/index.php?siteid=238202), se puede que los nombres presentados en la BD no son los mismos a los indicados en el ECGH, por lo que es importante el conocimiento del tema para determinar cada uno de los elementos que contiene la base de datos. En la Ilustración 39, se aprecia por colores la información válida, que en este caso es Strand (hélice del ADN) en color verde, id del gen color rojo, id de transcripción color amarillo y nombre de la proteína color rosado.

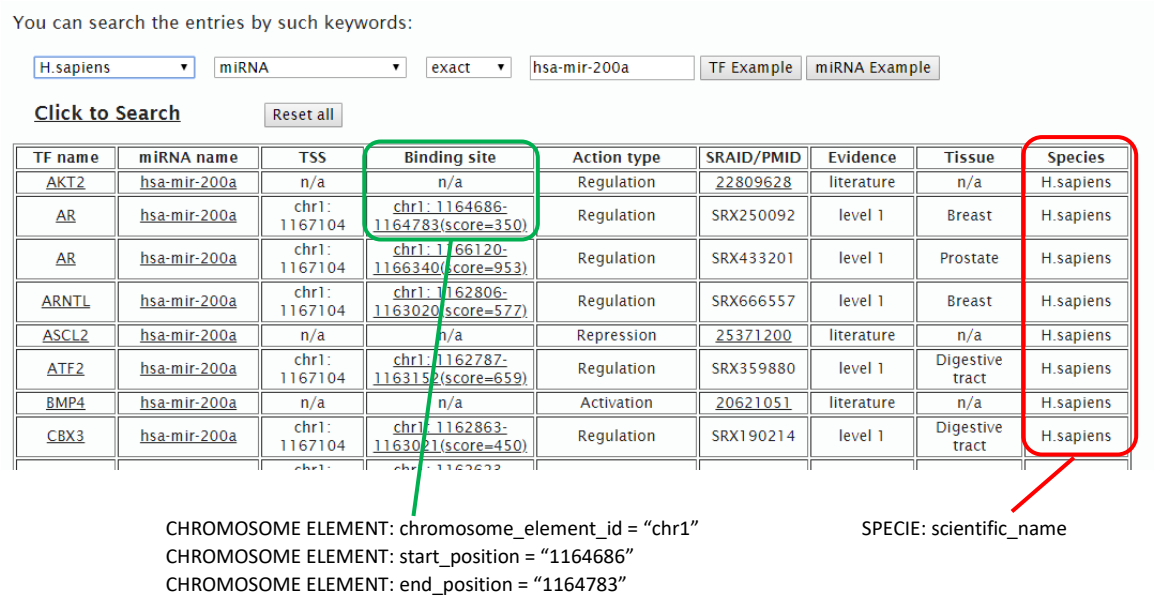
Ilustración 39 Ejemplo problema 2



Fuente: Propia

EJEMPLO 3: En la base de datos genómica TransmiR (http://www.cuilab.cn/transmir), se puede ver en la Ilustración 40 que el campo "binding site" se refiere a la localización (Cromosoma, inicio, fin) color verde, la interpretación de "Specie" es representado en el ECGH como "scientific_name" pero en la mayoría de bases de datos el nombre científico es Homo Sapiens, en este caso se lo encuentra abreviado como "H. sapiens".

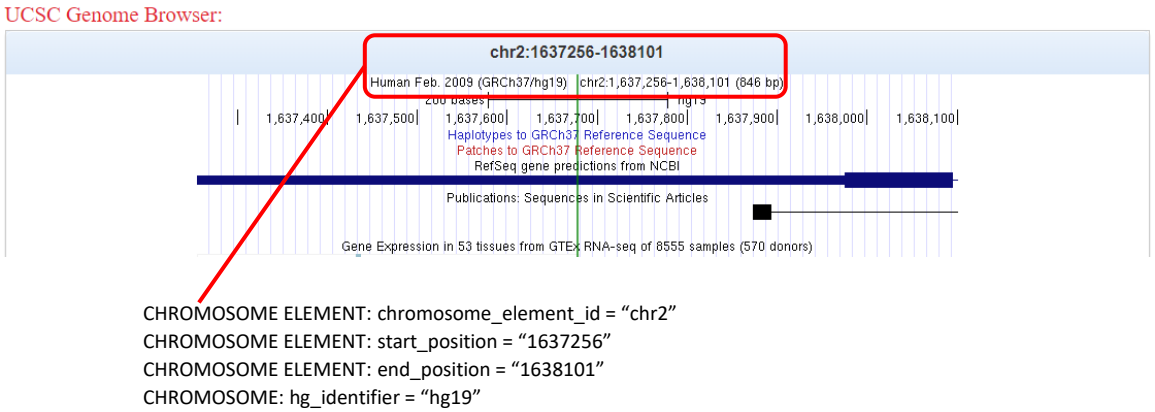
Ilustración 40 Ejemplo problema 3



Fuente: Propia

EJEMPLO 4: Tomado de la base de datos "USCG Genome Browser", en este caso se tiene que la presentación de la información se la da en base a gráficos. En la Ilustración 41 se puede observar que la localización del gen si bien se la indica de manera completa pues es complicado poder realizar la extracción de dicha información mediante un script al ser una imagen la visualizada.

Ilustración 41 Ejemplo problema 4



Fuente: Propia

EJEMPLO 5: Tomando la base de datos genómica "Primary Carnitine Deficiency (OCTN2) and SLC22A5 gene Database" (http://www.arup.utah.edu/database/OCTN2/OCTN2_display.php), se puede apreciar en la Ilustración 42 que la columna "Classification" forma parte del ECGH bajo el nombre del atributo "clinically_important" esto en color rojo. Adicionalmente se indica que hay información sobre mutaciones (color verde) así que en el ECGH pertenece a "Mutation" y en base a lo que se indique en este campo se tomará la información de la columna "Nucleotide Change". Esta columna permite obtener los alelos que han sido modificados debido a la mutación siendo en el ECGH los atributos SNP_GENOTYPE: allele1 y allele2.

Ilustración 42 Ejemplo problema 5

271 variants found

Search:

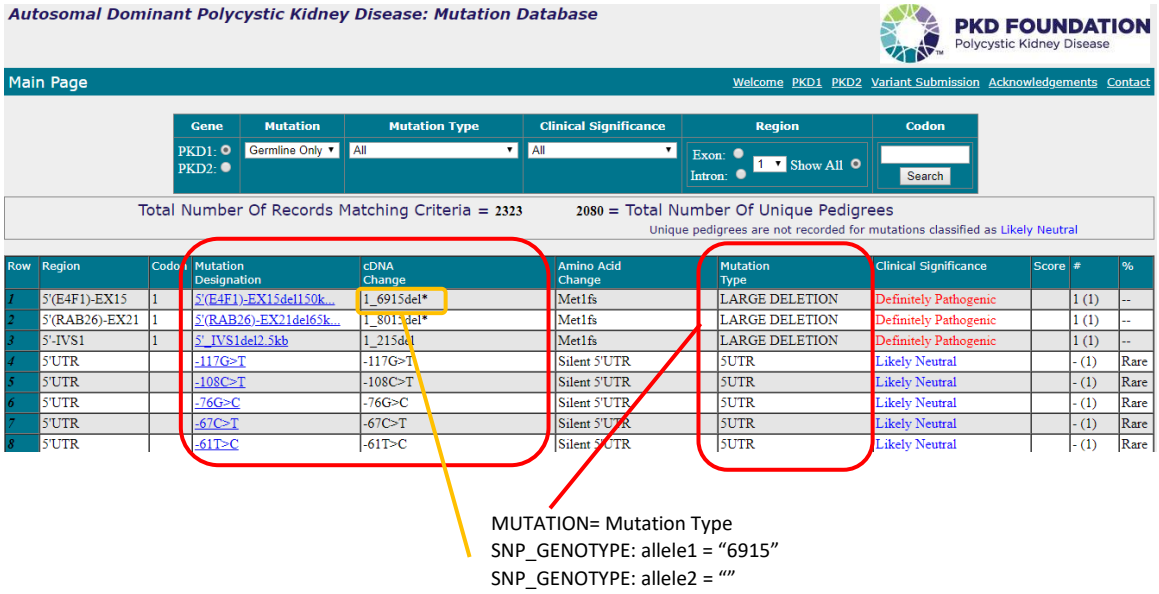
Location	Mutation Type	Nucleotide Change	Protein Change	Transport Activity	Expression	Classification	References	Comments
	Missense		p.G152D			VUS	Calderon et.al unpublished	
Exon 3	Missense	c.631T>C	p.Y211H			VUS	Calderon et.al unpublished	
5'UTR		c.-207G>C				Benign	Peltekova (2004)	
5'UTR		c.-185A>C		33		Uncertain	Calderon et.al unpublished	
5'UTR		c.-149G>A		33		Uncertain	Calderon et.al unpublished	
5'UTR	Deletion	c.-91_22del				Pathogenic	Nezu (1999)	
5'UTR		c.-78C>T		33		Benign	Koizumi (1999)	
5'UTR		c.-77G>A		33		Benign	Koizumi (1999)	
5'UTR		c.-38A>C				Benign	Calderon et.al unpublished	
Exon 1	Missense	c.3G>T	p.M1I	<5		Likely Pathogenic	Dobrowolski (2005)	

DELETION: bases = "c"
SNP_GENOTYPE: allele1 = "91"
SNP_GENOTYPE: allele2 = "22"

Fuente: Propia

EJEMPLO 6: Tomando de la base de datos genómica "Autosomal Dominant Polycystic Kidney Disease" (http://pkdb.mayo.edu/cgi-bin/v2_display_mutations.cgi?apkd_mode=PROD). La Ilustración 43 indica el tipo de mutación correspondiente en el ECGH a "MUTATION" de manera similar se considera en base a este campo la información que presenta la columna "cDNA Change" ya que en ella se indica la operación que se realizó.

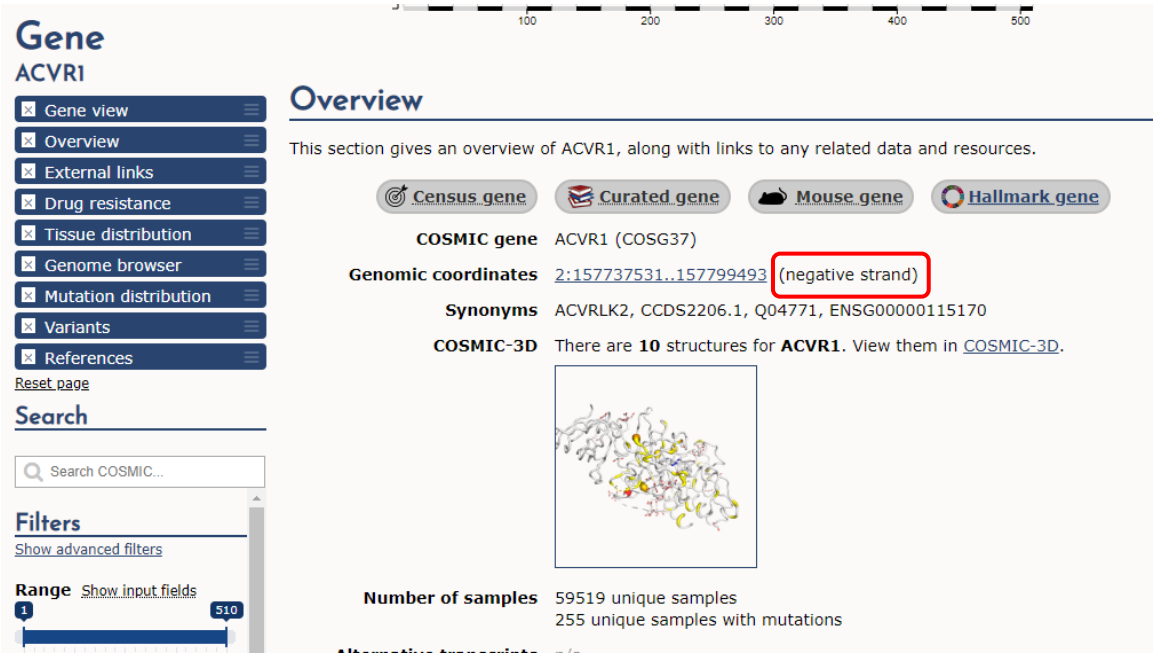
Ilustración 43 Ejemplo problema 6



Fuente: Propia

EJEMPLO 7: Tomando la base de datos genómica de "ACVR1" (https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=ACVR1) se puede apreciar en la Ilustración 44 que lo enmarcado en color rojo pertenece en el ECGH a Strand (hélice del ADN), pero en este caso se encuentra indicado como "negatice strand".

Ilustración 44 Ejemplo problema 7



Fuente: Propia

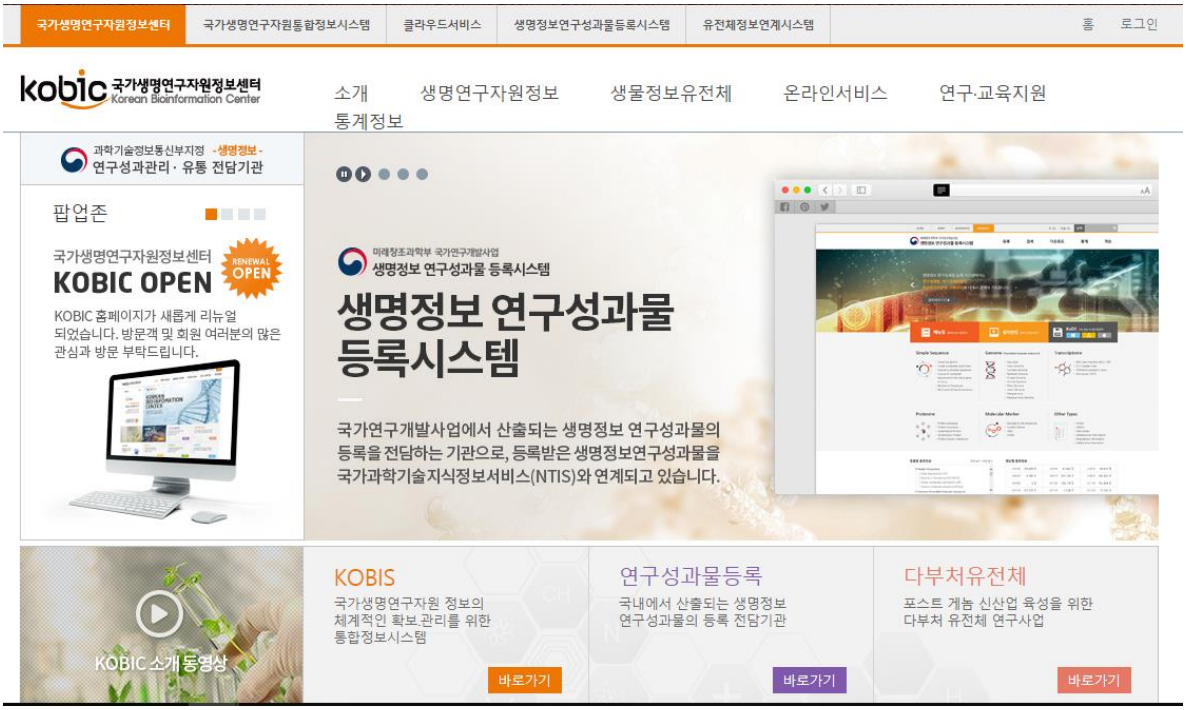
EJEMPLO 8: Considerando la base de datos genómica "Comparasite" (http://www.ims.u-tokyo.ac.jp/imsut/jp/), se puede apreciar (Ver Ilustración 45) que se encuentra en otro idioma, pero el contenido también se lo puede visualizar en Inglés. El desconocimiento del idioma puede llevar a confusión de la información que se extrae, en este caso se puede obtener traducción propia del sitio web.

Ilustración 45 Ejemplo problema 8

Fuente: Propia

EJEMPLO 9: El contenido presentado en la base de datos "CleanEST" (https://www.kobic.re.kr/) se encuentra en un idioma diferente al inglés en el que usualmente se encuentra todo el material genético. En el ejemplo 8 se disponía de la traducción propia del sitio, pero en este caso no se obtiene traducción alguna (Ver Ilustración 46), por lo que se genera una controversia en el contenido que el sitio web presenta.

Ilustración 46 Ejemplo problema 9



Fuente: Propia

- Varias de las bases de datos genómicas indican quienes han sido sus curators [1] o revisores de la información publicada, no obstante, en algunos de los sitios no se indica ni refleja dicha información. Para mitigar el caso de que no se indique el personal que revisó y validó la información presentada se consideró agregar el nombre del sitio web oficial quien se hace cargo de todo lo publicado.

- Otro de los inconvenientes encontrados es que diferentes sitios web no han recibido mantenimiento. En algunos casos los servidores han dejado de funcionar y algunas páginas web han perdido incluso su dominio por lo que no se obtiene ni siquiera una portada o datos informativos sobre el proyecto que trabajaban. Este es un punto clave ya que algunos sitios pudiesen dar información valiosa pero al perderse el rastro simplemente se omiten e ignora la base de datos genómica.

  Algunos sitios web en donde las bases de datos genómicas eran públicas en un inicio, ahora se han convertido en sitios de pago o a su vez se requiere de una invitación para la creación de una cuenta de acceso. Es así que para comprobación de los recursos muchos de ellos pueden encontrarse restringidos.

5.2 CONCLUSIONES

Una vez concluido el trabajo de exploración de bases de datos genómicas dirigida por modelos conceptuales se ha llegado a las siguientes conclusiones:

- La exploración de cada una de las bases de datos genómicas permitió especificar la correspondencia entre los diferentes atributos requeridos por el ECGH. El mapeo de cada elemento generó una trazabilidad entre el ECGH y las DB genómicas, de esta manera se identificó las diferentes áreas de conocimiento que se manejan en el ámbito genómico.

- La utilización de un esquema conceptual permite tener una versatilidad en el manejo de la información en este caso enfocado al genoma humano. Esto se debe a que los atributos requeridos por investigadores o médicos son adaptados a medida de los nuevos conocimientos adquiridos en base a la investigación científica. Si se requiere incorporar nueva información, el modelo conceptual permite realizarlo sin inconvenientes, al igual que si se requiere omitir o eliminar.

- Se identificó los problemas comunes a la hora de integrar información heterogénea. Las distintas bases de datos presentan información poco estandarizada. Si bien los sitios web utilizan nomenclaturas del área genómica, se torna complicado realizar un mapeo entre lo presentado en el sitio web con lo extraído en un ECGH, esto se debe a que la información no se encuentra en los formatos deseados o está incompleta.

[1] ***Curator***: Persona que verifica y valida la información antes de ser publicada. Se hace responsable por los datos presentados como resultado de una investigación.

- La base datos que se ha conformado en este trabajo permite tener una clara visión de las bases de datos que se encuentran activas y que tipo de información contiene. A su vez el resultado de este trabajo sienta una base para la construcción de un SIGe, en el cual los investigadores o expertos en el ámbito genético puedan analizar y explotar la información almacenada con fines de diagnósticos.

- La automatización para la exploración y extracción de las bases de datos genómicas es aún ambigua, ya que en su gran mayoría los sitios web contienen información escueta. El porcentaje de sitios que permiten la utilización de servicios web, APIs u otros mecanismos de acceso imposibilitan el manejo de la información e integración con otros sistemas. Para la gran mayoría de casos la revisión de los recursos aún debe realizarse de manera manual. No obstante, para los sitios en donde se posee una API, se pueden generar alternativas para la extracción y gestión mediante scripts.

5.3 TRABAJOS FUTUROS

Al concluir este trabajo se ha proyectado diferentes trabajos que pudiesen realizarse a futuro, entre ellos se destaca:

- La incorporación de algoritmos o scripts que permitan constantemente monitorear en base al mapeo realizado actualmente, en donde se verifique:

    - Cambios en los listados fuente, ya que se pudieron incorporar nuevas bases de datos.
    - Las bases de datos actualmente recabadas pueden ser obsoletas.
    - Las bases de datos pudieron incorporar nuevos campos en la presentación de resultados.

- Revisión y adaptación de cambios en los datos del esquema conceptual del genoma humano, ya que con cada uno de los diferentes estudios y descubrimientos que surgen constantemente, la información u orientación de los datos bases del ECGH pueden verse comprometidos. Al realizar una revisión el ECGH se actualizaría e incorporaría información que inicialmente no ha sido considerada.

- Realizar una ampliación en el número de las bases de datos, considerando los links externos que pueden presentarse en los sitios web. En varias ocasiones los sitios web de las bases de datos genómicas presentan información asociada que pertenece o reposa en otros lugares, por lo que se puede expandir la información que se maneja y agregar nuevas bases de datos genómicas referentes al genoma humano.

- Generar un sistema recomendador de bases de datos genómicas en los que se puedan incluir cuadros de mando, en donde se pueda apreciar porcentajes de completitud de información, interfaces gráficas de exploración, entre otros.

REFERENCIAS BIBLIOGRÁFICAS

[1] C. E. Cook, M. T. C. G. Bergman, R. Apweiler y E. Birney, «The European Bioinformatics Institute in 2017: data coordination and integration,» *Nucleic Acids Research,* vol. 46, p. D21–D29, 4 Enero 2017.

[2] UK government, «Strategy for UK life sceinces: one year on,» 2012. [En línea]. Available: https://www.gov.uk/government/publications/strategy-for-uk-life-sciences-one-year-on.

[3] R. Mirnezami, J. Nicholson y A. Darzi, «Preparing for precision medicine,» *New England Journal of Medicine,* vol. 6, nº 366, pp. 489-491, 2012.

[4] A. Middleton, «Society and personal genome data,» *Human Molecular Genetics,* vol. 27, nº Issue R1, p. R8–R13, 2018.

[5] A. M., d. S. J. J., P. C., J. R. y C.-B. S., «Public responses to the sharing and linkage of health data for research purposes: a systematic review and thematic synthesis of qualitative studies,» *BMC Med. Ethics,* vol. 17, p. 73, 2016.

[6] G. D., M. N., S. A., W. F. y A. D.A., «Public preferences about secondary uses of electronic health information,» *JAMA Int. Med,* nº 173, pp. 1798-1806, 2013.

[7] O. A., Conceptual Modeling of Information Systems, 1 ed., Berlin: Heidelberg: Springer Berlin Heidelberg, 2007.

[8] D. Aguilera, C. Gómez y A. Olivé, Enforcement of Conceptual Schema Quality Issues in Current Integrated Development Environments, 2013, p. 626–640.

[9] N. W. Paton y e. al., Conceptual modelling of genomic information, 6 ed., vol. 16, 2000, p. 548–557.

[10] N. W. Paton y E. Bornberg-bauer, Conceptual data modelling for bioinformatics, 2 ed., vol. 3, 2002, p. 166–180.

[11] S. Ram y W. Wei, Modeling the Semantics of 3D Protein Structures, in Genome, 2004, p. 696–708.

[12] A. Bernasconi, S. Ceri, A. Campi y M. Masseroli, «Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data,» de *In International Conference on Conceptual Modeling*, Springer, Cham, 2017.

[13] R. Wieringa, «Design science methodology for information systems and software engineering,» *Springer Verlag,* 2014.

[14] J. C. L. &. N. B. Mylopoulos, «Representing and using nonfunctional requirements: A process-oriented approach.,» *IEEE Transactions on software engineering,* vol. 18, nº 6, pp. 483-497, 1992.

[15] B. Thalheim, The theory of conceptual models, the theory of conceptual modelling and foundations of conceptual modelling., Berlin, Heidelberg: Springer, 2011, pp. 543-577.

[16] C. G. M.-R. S. a. E. T. Jordi Cabot, «30 Years of Contributions to Conceptual,» de *Conceptual Modeling Perperctives.*, Springer, 2017, pp. 7-20.

[17] G. Muller, System and context modeling - the role of time-boxing and multi-view interaction., vol. 3, Syst. Res. Forum, 2009, pp. 139-152.

[18] J. Ludewig, «Models in software engineering–an introduction.,» *Software and Systems Modeling,* vol. 2, nº 1, pp. 5-14, 2003.

[19] Ó. Pastor, A. León, J. Reyes y J. C. Casamayor, «Modeling Life: A Conceptual Schema-centric Approach to Understand the Genome,» de *Conceptual Modeling Perspectives*, Springer, 2017, pp. 25-38.

[20] J. F. Reyes Román, Diseño y Desarrollo de un Sistema de Información Genómica Basado en un Modelo Conceptual Holístico del Genoma Humano (Doctoral dissertation), Valencia: Tesis Doctoral, 2018.

[21] J. Xu, Next-generation Sequencing, Caister : Academic Press, 2014.

[22] Oxford University, «NAR Database Summary Paper,» Oxford University Press, 2014. [En línea]. Available: www.oxfordjournals.org.

[23] National Center for Biotechnology Information, «All resources,» National Center for Biotechnology Information, [En línea]. Available: https://www.ncbi.nlm.nih.gov/.

[24] Human Genome Variation Society, «Databases & Tools,» HGVS, [En línea]. Available: http://www.hgvs.org/content/databases-tools.

[25] University of Pittsburgh, «search.HSLS.OBRC,» The Health Sciences Library System, 2014. [En línea]. Available: https://www.hsls.pitt.edu/obrc/.

[26] MedicineNet Inc., «Medical Definition of Genome Database,» MedicineNet, 24 01 2017. [En línea]. Available: https://www.medicinenet.com/script/main/art.asp?articlekey=10721.

[27] T. K. Attwood, A. Gisel, N. E. Eriksson y E. Bongcam-Rudloff, «Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective.,» *Bioinformatics-trends and methodologies,* vol. Tech, 2011.

[28] P. Bourne, «Will a Biological Database Be Different from a Biological Journal?,» *PLoS Comput Biol,* vol. 1, nº 3, pp. 179-181, 2005.

[29] Health Sciences Library Systems, «OBRC: Online Bioinformatics Resources Collections,» University of Pittsburgh, 2014. [En línea]. Available: https://www.hsls.pitt.edu/obrc/.

[30] Oxford University Press, «Nucleic Acids Research,» Oxford University Press, 2018. [En línea]. Available: https://academic.oup.com/nar/pages/About.

[31] Oxford University Press, «NAR Database Summary Paper Category List,» Oxford University Press, 2014. [En línea]. Available: https://www.oxfordjournals.org/nar/database/c/.

[32] J. Ostell, «What's in a Genome at NCBI?,» National Center for Biotechnology Information (US), 8 Noviembre 2013. [En línea]. Available: https://www.ncbi.nlm.nih.gov/books/NBK169442/.

[33] A. R. B. T. B. J. B. D. B. C. B. E. B. S. C. K. C. D. C. K. D. M. D. I. F. S. F. M. G. L. G. V. H. M. J. M. K. C. K. V. K. A. K. M. Acland A, «Database resources of the National Center for Biotechnology Information,» *Nucleic acids research,* nº 41, p. D8–D20, 2013.

[34] A. R. B. A. C. S. C. S. F. B. e. a. Klimke W, «The National Center for Biotechnology Information's Protein Clusters Database,» *Nucleic acids research,* nº 37, p. D216–23, 2009.

[35] K. K. S. H. M. D. Pruitt KD, «Introducing RefSeq and LocusLink: curated human genome resources at the NCBI,» *Trends Genet,* vol. 16, nº 1, p. 44–47, 2000.

[36] HGVS, «About HGVS,» HGVS, 01 Julio 2018. [En línea]. Available: http://www.hgvs.org/content/about-hgvs.

[37] HGVS, «Recommendations for the description of sequence variants,» HGVS, 22 Marzo 2016. [En línea]. Available: http://www.hgvs.org/mutnomen/recs.html.

[38] HGVS, «History regarding the description of sequence variants,» HGVS, 21 Diciembre 2015. [En línea]. Available: http://www.hgvs.org/mutnomen/history.html.

[39] Health Sciences Library System, University of Pittsburgh, «About HSLS,» 2018. [En línea]. Available: https://www.hsls.pitt.edu/about-us.

[40] Department of Health & Human, «Health Sciences Library System (HSLS),» UNIVERSITY of PITTSBURGH, 2016. [En línea]. Available: http://files.hsls.pitt.edu/files/hsls-overview.pdf.

[41] University of Pittsburgh, «Research Support,» Health Sciences Library System, 2018. [En línea]. Available: https://www.hsls.pitt.edu/research-support.

[42] University of Pittsburgh, «Consultation,» Health Sciences Library System, 2018. [En línea]. Available: https://www.hsls.pitt.edu/consultation.

[43] Python Software Foundation, «io — Core tools for working with streams,» Python Software Foundation, 2018. [En línea]. Available: https://docs.python.org/3/library/io.html.

[44] Python Software Foundation, «csv — CSV File Reading and Writing,» Python Software Foundation, 2018. [En línea]. Available: https://docs.python.org/3/library/csv.html.

[45] Python Software Foundation, «httplib — HTTP protocol client,» Python Software Foundation, 21 Junio 2018. [En línea]. Available: https://docs.python.org/2/library/httplib.html.

[46] Python Software Foundation, «urllib.parse — Parse URLs into components,» Python Software Foundation, 2018. [En línea]. Available: https://docs.python.org/3/library/urllib.parse.html.

[47] Clinvar, «National Center for Biotechnology Information,» 2016. [En línea]. Available: www.nbci.nlm.nih.gov/clinvar/.

[48] Ensambl, «About the Ensembl Project,» 2018. [En línea]. Available: www.ensembl.org.

[49] G. L. Michael Cariaso, «SNPedia: a wiki supporting personal genome annotation, interpretation and analysis, Nucleic Acids Research,» vol. 40, p. D1308–D1312, 2012.

[50] M. Krötzsch, D. Vrandečić, M. Völkel, H. Haller y R. Studer, «Semantic wikipedia.,» *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 5, nº 4, pp. 251-261, 2007.

[51] M. Ashburner y e. al., Gene Ontology: tool for the unification of biology, 1 ed., vol. 25, Nat. Genet., 2000, p. 25–29.

ANEXOS

ANEXO 1 – ESQUEMA CONCEPTUAL DEL GENOMA HUMANO

La información de este anexo ha sido tomada de la Tesis doctoral de José Reyes [20].

VISTA ESTRUCTURAL

A continuación, se presenta un detalle del contenido de las clases que pertenecen a esta vista:

CLASE HOTSPOT

Asocia todos los valores correspondientes a los puntos en la secuencia de ADN donde puede existir una posible recombinación en la meiosis. Tiene los siguientes atributos:

- ***hotspot_id*** [string]: Identificador del cruce de recombinación
- ***position*** [string]: Punto en la secuencia de ADN donde se produce la recombinación.

CLASE CYTOBAND

Contienen la información de las subregiones del cromosoma conocido como banda citogenética. Tiene como atributos:

- ***name*** [string]: Puede ser una "q" o "p" según el brazo del cromosoma, seguido de números que dependerán de la resolución utilizada.
- ***score*** [string]: Representa la intensidad del tintado.
- ***start_position*** [long]: Posición inicial en la secuencia de referencia del cromosoma.
- ***end_position*** [long]: Posición final en la secuencia de referencia del cromosoma.

CLASE SPECIES

Es el grupo de organismos que poseen un ADN semejante, permite la determinación de la especie a la que forma parte el ADN, esta puede ser: humana, animal, vegetal, entre otras. Tiene como atributos:

- ***scientific_name <<oid>>*** [short]: Identificador de la especie.
- ***common_name*** [string]: Nombre de la especie
- ***nbci_taxon_id*** [string]: Identificador de la especie otorgado por la NBCI.
- ***assembly*** [string]: Identificador de la versión de secuencia genómica.
- ***date_assembly*** [string]: Fecha de la versión genómica.
- ***source*** [string]: Fuente de donde se obtiene la secuencia genómica.

CLASE CHROMOSOME

Contiene la información genética ya sea de genes, elementos reguladores y tras secuencias de nucleótidos. Tiene como atributos:

- ***nc_identifier<<oid>>*** [string]: Identificador interno de la secuencia cromosómica, es proporcionado por NCBI.
- ***name*** [string]: Nombre del cromosoma en algunas bases de datos se lo considera simplemente como un número que es tomado del orden que se encuentra en el ADN.
- ***hg_identifier*** [string]: Identificador de la versión del genoma utilizado.
- ***sequence*** [long]: Secuencia de referencia del cromosoma.

CLASE CHROMOSOME ELEMENT

Representa los fragmentos del cromosoma con alguna significación. Tiene los siguientes atributos:

- ***chromosome_element_id*** [string]: Identificador interno de cada elemento del cromosoma.
- ***start_position*** [long]: Posición inicial del elemento.
- ***end_position*** [long]: Posición final del elemento.
- ***strand*** [string]: Indica la hebra (hebra positiva o negativa de la hélice) en la que se encuentra el elemento.
- ***specialization_type*** [string]: Indica si es un elemento de transcripción, regulación o conservado.

VISTA DE TRANSCRIPCIÓN

A continuación, se detalla las clases que la conforman:

CLASE TRANSCRIPTABLE ELEMENT

Representa a los elementos que crean un ARN complementario partiendo de una secuencia de ADN siendo las especializaciones los genes y exones.

CLASE GENE

Representa la información de una partícula de material genético que determinan las diferentes características hereditarias de las especies. Tiene los siguientes atributos:

- ***id_symbol*** [string]: Identificador de un gen mediante una abreviatura.
- ***id_hugo*** [string]: Identificador otorgado por el consorcio HGNC[2] .
- ***oficial_name*** [string]: Nombre oficial completo de un gen.
- ***description*** [string]: Descripción del gen.
- ***biotype*** [string]: Indica el tipo de gen.
- ***status*** [string]: Estado de validez del elemento.
- ***gc_percentage*** [float]: Porcentaje de contenido de Gs y Cs en la secuencia.
- ***gene_synonym*** [string]: Sinónimo con el que también se le conoce a un gen.
- ***start_geneng*** [int]: Posición inicial del gen dentro de la secuencia.
- ***end_geneng*** [int]: Posición final del gen dentro de la secuencia.

CLASE TF

Esta clase representa a los genes que codifican una proteína. Tiene el siguiente atributo:

- ***cons_seq*** [string]: Secuencia de nucleótidos.

CLASE miRNA

Representa a los RNA no codificantes con respecto a una proteína.

CLASE EXON

Representa a los elementos transcribibles que forman parte del gen y que son unidades básicas de los transcritos. Tiene los siguientes atributos:

- ***name*** [string]: Nombre del exón proporcionado por NCBI a cada uno de los exones.

[2] **HGNC**: Gene Nomenclature Committee

- ***id_symbol*** [string]: Clave que indica a qué gen pertenece el exón.
- ***start_exonng*** [int]: Posición de inicio del exón en la secuencia de referencia del gen.
- ***end_exonng*** [int]: Posición donde finaliza el exón en la secuencia de referencia del gen.

CLASE REGULATORY ELEMENT
Representa las regiones del ADN que realizan una función reguladora controlando algunos procesos existentes en el ADN.

CLASE GENE REGULATOR
Representa los elementos reguladores del gen.

CLASE TFBS
Representa a las regiones de unión de los factores de transcripción que producen algún efecto en la transcripción del gen, ya sea de activación o represión. Tiene los siguientes atributos:

- ***name*** [long]: Nombre del sitio de unión de factores de transcripción.
- ***type*** [string]: Sitios de unión de los factores de transcripción.
- ***description*** [string]: Descripción de los sitios de unión del factor de transcripción.
- ***score*** [float]: Grado de similitud entre la secuencia consenso y el tfbs.
- ***cons_seq*** [string]: Secuencia consenso la cual enlaza el tfbs.

CLASE CPG ISLAND
Representa las regiones en las que existe una concentración de pares de Cs y Gs enlazados por fosfatos. Tiene el siguiente atributo:

- ***cp_percentage*** [float]: Porcentaje de GC en el elemento.

CLASE TRIPLEX
Representa a las secuencias de ADN que se intercalan en la doble hélice de ADN de las células.

CLASE PROMOTER
Representa la región de ADN que controla la iniciación de la transcripción de una determinada parte del ADN a ARN.

CLASE TRANSCRIPT REGULATOR
Representa las regiones reguladoras del transcrito.

CLASE MIRNA TARGET
Representa la región reguladora del transcrito a la que se unirá post-transcripcionalmente un miRNA.

CLASE SPLICING REGULADOR
Representa un elemento regulador del transcrito que regula el proceso de splicing. Tiene los siguientes atributos:

- ***type*** [String]: Tipo de regulación.
- ***regulated_element*** [String]: Elemento regulado siendo un intrón o exón.

CLASE CONSERVED REGION

Representa las regiones conservadas dentro del cromosoma, es decir las regiones que no se han modificado ni cambiado. Tiene el siguiente atributo:

- ***score*** [float]: Grado de conservación de la región.

CLASE TRANSCRIPT

Representa los diferentes transcritos que presenta un gen. Tiene los siguientes atributos:

- ***transcript_id*** [int]: Identificador interno del transcrito
- ***biotype*** [string]: Indica el tipo al que pertenece.
- ***startcds*** [int]: Posición inicial del proceso de traducción a proteína.
- ***endcds*** [int]: Posición final del proceso de traducción.
- ***nm_identifier*** [string]: Identificador proporcionado por NCBI.
- ***start_transcriptng*** [int]: Inicio del transcrito dentro de la secuencia del gen.
- ***end_transcriptng*** [int]: Fin del transcrito en la secuencia del gen.

CLASE PROTEIN CODING

Al ser una clase de especialización esta representa a las proteínas que han sido codificadas. Tiene los siguientes atributos:

- ***start_position_ORF*** [int]: Posición inicial de la secuencia codificada.
- ***end_position_ORF*** [int]: Posición final de la secuencia codificada.

CLASE PROTEIN

Representa a las proteínas que han sido sintetizadas a partir de un transcrito. Tiene los siguientes atributos:

- ***chr_transcript_id*** [int]: Identificador interno del transcrito.
- ***name*** [string]: Nombre de la proteína.
- ***sequence*** [long]: Secuencia de la proteína.
- ***source*** [string]: Fuente de datos de donde se extrae la información.
- ***np_identifier*** [string]: Identificador interno proporcionado por NCBI.
- ***accession*** [string]: Identificador que indica la fuente de datos de donde se extrajo la información.

VISTA DE VARIACIONES

A continuación, se detalla cada uno de los atributos que contienen:

VARIATION

Representa a cada una de las variaciones que se encuentran en el ADN. Tiene los siguientes atributos:

- ***variation_id<<oid>>*** [int]: Identificador de la variación.
- ***chr_gene_id*** [int]: Clave foránea que permite la vinculación con gen.
- ***db_version_id*** [int]: Versión y base de datos fuente de la variación.
- ***description*** [string]: Descripción de la variación.
- ***db_vatiation_id*** [string]: Identificador de la fuente de la variación.

- ***clinically_important*** [string]: Descripción de la importancia clínica sobre la variación.
- ***private*** [int]: Indica mediante 1 o 0 a manera de booleano si la variación es privada o no.
- ***nc_identifier*** [string]: Identificador de la versión de secuencia cromosómica.
- ***ng_identifier*** [string]: Identificador de la versión de secuencia génica.
- ***other_identifier*** [string]: Posibles identificadores que presenten otras fuentes de datos.
- ***associated_genes*** [string]: Genes asociados a la variación.
- ***omim*** [string]: Versión de OMIM[3] para la variación.
- ***creation_version*** [string]: Versión de la variación al momento de ser creada.

CLASE MUTATION

Esta clase hace referencia a las variaciones con bajo efecto patológico (<1%) en una población.

CLASE POLIMORPHISM

Esta clase describe las variaciones con porcentaje mayor a 1% en una población.

CLASE CNV

La clase referida como número de copia variación define las repeticiones o número de borrado de una región de la secuencia de ADN.

CLASE SNP

SNP es un polimorfismo particular que se genera cuando un único nucleótido dentro del genoma difiere de lo habitual entre individuos de la misma especie que ha sido agrupada por poblaciones. Tiene el siguiente atributo:

- ***map_weight*** [int]: Número de veces que se ha mapeado el SNP en el genoma.

SNP_ALLELE

Es la representación de los distintos valores que toma un SNP con relación a un solo alelo. Tiene el siguiente atributo:

- ***allele*** [char]: Valor que tiene el alelo (A, T, G, C).

CLASE SNP_ALLELE_POP

Representa la frecuencia en la que cada SNP para un alelo aparece en cada población. Tiene el siguiente atributo:

- ***frequency*** [float]: Frecuencia de aparición en las diferentes poblaciones.

CLASE SNP_GENOTYPE

Representa los valores que toma un par de alelos cada individuo en la posición del SNP para las dos hebras. Tiene los siguientes atributos:

- ***allele1*** [string]: Valor que toma un alelo en una hebra.
- ***Allele2*** [string]: Valor que toma un alelo en una hebra.

[3] **OMIM**: Online Mendelian Inheritance in Man

CLASE SNP_GENOTYPE_POP

Representa la frecuencia en la que cada SNP aparece en la población con referencia a los dos alelos. Tiene el siguiente atributo:

- ***frequency*** [float]: Frecuencia de aparición en las diferentes poblaciones.

CLASE POPULATION

Representa las poblaciones o conjuntos de individuos con características comunes. Tiene los siguientes atributos:

- ***id*** [int]: Identificador interno de la población.
- ***name*** [string]: Nombre de cada población.
- ***description*** [string]: Descripción de cada población.
- ***size*** [int]: Número de individuos pertenecientes a la población.

CLASE LD

Representa a la relación entre dos SNPs en una población específica. Tiene los siguientes atributos:

- ***Dprime*** [double]:
- ***Rsqueare*** [double]:
- ***LOD*** [double]:

POP_ORIGIN_DB

Representa a la base de datos de la cual se ha extraído la población. Tiene los siguientes atributos:

- ***pop_orig_id*** [int]: Identificador de la base de datos.
- ***name*** [string]: Nombre de la base de datos.
- ***description*** [string]: Descripción de la población a la que se hace referencia.
- ***URL_pop*** [string]: Dirección web de la base de datos que proporciona la información de la población.

CLASE PRECISE

Representa las variaciones que han sido detectadas su posición en el cromosoma dentro de la secuencia del ADN. Tiene los siguientes atributos:

- ***specialization_type*** [string]: Tipo de variación dentro del genoma.
- ***flanking_right*** [string]: Secuencia de 20 nucleótidos a la derecha.
- ***flanking_left*** [string]: Secuencia de 20 nucleótidos a la izquierda.
- ***aln_quality*** [int]: validad del alineamiento dentro del gen.
- ***position*** [int]: Posición en la que se encuentra la variación dentro de la secuencia del cromosoma.

CLASE INSERTION

Representa la adición de una o más secuencias de nucleótidos un determinado número de veces en una secuencia de ADN. Tiene los siguientes atributos:

- ***sequence*** [string]: Secuencia de nucleótidos insertados en la secuencia.
- ***repetition*** [int]: Número de veces que se repite la secuencia insertada.

CLASE DELETION
Representa la eliminación o borrado de un número de nucleótidos en la secuencia de ADN. Tiene el siguiente atributo:

- ***bases*** [int]: Número de nucleótidos borrados en la secuencia.

CLASE INDEL
Representa las operaciones de inserción y eliminación que ocurren de manera simultánea, es decir el reemplazamiento de nucleótidos en la secuencia de ADN. Tiene los siguientes atributos:

- ***ins_sequence*** [string]: Secuencia de nucleótidos insertados en la secuencia de ADN.
- ***ins_repetition*** [int]: Número de veces que se repite la secuencia insertada.
- ***del_base*** [int]: Número de nucleótidos borrados o eliminados de la secuencia de ADN

CLASE INVERSION
Representa el cambio de orden de una secuencia de nucleótidos dentro de la secuencia de ADN. Tiene el siguiente atributo:

- ***bases*** [int]: Número de nucleótidos que han sido invertidos dentro de la secuencia.

CLASE IMPRECISE
Representa la variación de la cual se desconoce la posición dentro de la secuencia del ADN. Tiene el siguiente atributo:

- ***description*** [string]: Descripción de la variación en leguaje natural.

CLASE PHENOTYPE
Representa los fenotipos asociados a una o diferentes variaciones del ADN. Tiene los siguientes atributos:

- ***phenotype_id*** [int]: Identificador del fenotipo.
- ***name*** [string]: Nombre del fenotipo.

CLASE CERTAINLY
Representa los niveles de certeza que estén asociados a los fenotipos con respecto a una o diferentes variaciones. Tiene el siguiente atributo:

- ***level_certainty*** [string]: Nivel de certeza entre el fenotipo y la variación.

VISTA DE RUTAS METABÓLICAS O PATHWAYS

A continuación, se presenta el detalle de cada una de las clases que forman parte de esta vista:

CLASE EVENT
Representa las combinaciones de procesos existentes en el organismo. Tiene los siguientes atributos:

- ***event_id*** [string]: Identificador interno del evento.
- ***name*** [string]: Nombre del evento.

CLASE PROCESS
Representa a los procesos simples. Tiene el siguiente atributo:

- ***type*** [string]: Tipo de proceso que se lleva a cabo.

CLASE PATHWAY
Representa a los procesos complejos. Tiene el siguiente atributo:

- ***type*** [string]: Tipo de ruta metabólica o pathway ejecutado.

CLASE TAKES_PART
Representa la participación de una entidad en un proceso.

- ***notes*** [string]: Comentarios u observaciones sobre la relación entre las entidades que conforman un proceso.

CLASE INPUT
Representa a las entidades de entrada a un proceso. Tiene el siguiente atributo:

- ***stoichiometry*** [int]: Cantidad de la entidad que es utilizada en el proceso.

CLASE OUTPUT
Representa el resultado final del proceso. Tiene el siguiente atributo:

- ***stoichiometry*** [int]: Cantidad producida de la entidad saliente por el proceso.

CLASE REGULATOR
Representa a los procesos reguladores existentes en las partes intermedias de la reacción. Tiene el siguiente atributo:

- ***type*** [string]: Entidad que controla un proceso ya sea activándola o inhibiéndola.

CLASE CATALYSIS
Representa al proceso donde se aumenta o disminuye la velocidad de una reacción química. Tiene el siguiente atributo:

- ***EC_number*** [string]: Identificador de la reacción asignado por la “Enzyme Commision”.

CLASE ENZYME
Representa a las proteínas que catalizan reacciones químicas. Tiene el siguiente atributo:

- ***name*** [string]: Nombre de la encima acorde a la reacción producida.

CLASE ENTITY
Representa el tipo de entidades que pueden participar en un proceso de una ruta metabólica. Tiene los siguientes atributos:

- ***entity_id*** [string]: Identificador interno.
- ***name*** [string]: Nombre de la entidad.

CLASE COMPLEX
Representa a las entidades que han sido formadas por la combinación de otras entidades más simples. Tiene el siguiente atributo.

- ***detection_method*** [string]: Técnica utilizada en la formación de la entidad.

CLASE COMPONENT

Representa la manera en la que una entidad "Complex" está formada por sus entidades más simples. Tiene los siguientes atributos:

- ***stoichiometry*** [int]: Cantidad de participación de la entidad compleja.
- ***interation_type*** [string]: Descripción de formación mediante componentes.

CLASE POLYMER

Representa a las entidades que son generadas por alguna entidad. Tiene los siguientes atributos:

- ***min*** [int]: Repeticiones mínimas de la entidad.
- ***max*** [int]: Repeticiones máximas de la entidad.

CLASE SIMPLE

Representa las entidades más simples que se han formado en un proceso. De esta clase heredan las clases dna_e, rna_e, protein_e, aminoacid_e, nucleotide_e y basic_e.

CLASE ENTITYSET

Representa un conjunto de entidades que participan de manera habitual conjuntamente en algunos procesos.

VISTA DE FUENTES DE DATOS Y BIBLIOGRAFÍA

A continuación, se detalla cada una de las clases que forman parte de esta vista:

CLASE DATA BANK

Representa al a información de la fuente de datos de donde se tomaron los datos. Tiene los siguientes atributos:

- ***name*** [string]: Nombre de la fuente de datos.
- ***description*** [string]: Descripción de la fuente de datos.
- ***url*** [string]: Dirección web de la fuente de datos.

CLASE DATA BANK VERSION

Esta clase contiene información sobre la versión de la base o fuente de datos original. Tiene los siguientes atributos:

- ***release*** [string]: Versión de la Fuente de datos.
- ***name*** [string]: Nombre de la Fuente de datos utilizada.
- ***date*** [date]: Ultima fecha de actualización de la fuente de datos.

CLASE ELEMENT DATABANK

Representa la información de los elementos del cromosoma con la fuente de datos origen. Tiene los siguientes atributos:

- ***db_version_id*** [int]: Identificador que asocial la versión de la fuente de datos.
- ***source_identifier*** [string]: Identificador del element del cromosoma en el banco de datos.

CLASE DATA BANK ENTITY IDENTIFICATION

Esta clase asocial cada una de las entidades de la ruta metabólica con la fuente de datos origen.

- ***source_identification*** [string]: Identificador de la Fuente de datos de la ruta metabólica.

CLASE BIBLIOGRAPHY DB

Representa las diferentes fuentes de datos donde se encuentra información de las publicaciones científicas. Tiene los siguientes atributos:

- ***URL*** [string]: Dirección web donde se encuentra la base de datos que es fuente de la extracción de información de las publicaciones científicas.
- ***name_db*** [string]: Nombre de la base de datos fuente de la información de las publicaciones científicas.
- ***pubmed_id*** [int]: Identificador de la publicación en la base de datos PubMed.

CLASE BIBLIOGRAPHY REFERENCE

Representa a cada uno de los elementos propios de las publicaciones científicas. Tiene los siguientes atributos:

- ***bibliography_reference_id*** [int]: Identificador de la referencia bibligráfica.
- ***title*** [string]: Tpiitulo de la publicación científica.
- ***abstract*** [string]: Resumen de la publicación científica.
- ***publication*** [string]: Contenido de referencia de la publicación científica.
- ***authors*** [string]: Nombre de los diferentes autores de la publicación científica.
- ***date_pub*** [string]: Fecha de la publicación.