UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

“Diseño y modelo preliminar de una plataforma de integración de datos clínicos y genómicos: aplicaciones en Alzheimer y Cáncer de Mama”

Patricio Miguel Araneda García

ACTIVIDAD FORMATIVA EQUIVALENTE PARA OPTAR AL GRADO DE MAGISTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Rodrigo Assar Cuevas

2016

Indice

Indice de figuras

Indice de Tablas

Resumen

Durante los últimos años la medicina traslacional ha surgido como un enfoque potente para el estudio de enfermedades complejas, en que la idea fundamental es fortalecer la retroalimentación entre los estudios en ciencias básicas y la clínica para mejorar los diagnósticos y tratamientos de los pacientes. Accediendo a mayor información del paciente, en particular genómica, se busca definir de mejor manera el fenotipo de su enfermedad y con ello decidir su mejor tratamiento.

Sin embargo, la gran cantidad y heterogeneidad de los datos disponibles hace complejo el descubrimiento de información relevante (definir el fenotipo). Para abordar este problema es necesario desarrollar un sistema que permita integrar los estudios realizados a cada paciente y asociar sus resultados.

En este trabajo se propone implementar una plataforma (Datagenomed) constituida por un modelo de base de datos "híbrida" basado en PostgreSQL y almacenamiento JSON (NoSQL) y un conjunto de herramientas computacionales que permitan asociar la información clínica del paciente con la información genómica. Un software de gestión de datos que registre tanto información clínica (diagnóstica) como los resultados de secuenciación de ADN y que permita la búsqueda de información pertinente en repositorios biológicos, añadiendo reportes estadísticos basados en el software R.

La plataforma se adaptó a dos casos de estudio: i) información sobre Alzheimer basado en el proyecto Fondecyt No. 1140423 "Fisiopatología de la Apatía en la Enfermedad de Alzheimer: Un Estudio Experimental de Neuropsicología y Neuroimagen" (CA) liderado por la Dra. Andrea Slachevsky y ii) información de cáncer de mama del proyecto Fondef N. D11I1029 "Incorporación de la Secuenciación de Última Generación en el Cuidado de los Pacientes con Cáncer" (CC) proporcionado por la Dra. Katherine Marcelain.

Los datos clínicos provinieron de recolección de fichas clínicas hospitalarias, junto a datos demográficos (solo para CA). Los datos genómicos se obtuvieron del análisis de archivos Fastq

de muestras de sangre y/o tejido procesados mediante next-generation DNA sequencing (NGS) (CC).

Para adaptarse a la naturaleza disímil de los datos registrados, la información se almacenó en un nuevo sistema de bases de datos híbrido, permitiendo tanto datos clínicos estructurados como datos genómicos de tipo documental.

La implementación resultante cuenta con un sistema de filtrado y búsquedas de términos en bases bibliográficas e información genómica en bases de datos biológicas; Pubmed, RefSeqGene, MedGen, dbSNP, Clinvar, Cosmic, Gene pudiendo agregarse otros recursos según necesidad.

El objetivo de esta tesis es diseñar e implementar un conjunto de herramientas de software para permitir procesos de extracción, transformación y carga (ETL) de información sobre las bases de datos creadas y permitir consultas en línea mediante webservice. Dichos webservice se construyeron utilizando software open source y las mejores prácticas de diseño de interface, fuerte prototipado y técnicas de desarrollo xtreme programming.

El fin último es que la información resultante esté disponible remotamente vía una plataforma que pueda ser consultada utilizando webservice desde cualquier sistema de registro clínico asociado.

Como resultado se construyó una plataforma basada en tecnología web soportado sobre un motor de base de datos PostgreSQL utilizando Knime como herramienta para procesos de ETL.

Abstract

In recent years translational medicine has emerged as a powerful tool for the study of complex diseases approach, the fundamental idea is to strengthen the feedback between basic and clinical studies to improve diagnosis and treatment of patients. Accessing more information on the patient, particularly genomics, seeks to better define the phenotype of the disease and thus determine their best treatment.

However, due to the large amount of data and its heterogeneity the discovery of relevant information becomes complex (defining the phenotype). To address this problem it is necessary to develop a system that integrate studies and associate the patient outcomes.

In this thesis we propose to implement a platform (DataGenomed) consisting of a database model and a set of computational tools that allow to associate clinical information with genomic information of patients. The proposed data management software to record clinical information (diagnostic) and the results of DNA sequencing and allows the search for relevant information in biological repositories, adding statistical reports based on the software R.

The platform will tested two case studies: i) information on Alzheimer disease based on Fondecyt No. 1140423 project "Apathy Pathophysiology of Alzheimer's Disease: An Experimental Study of Neuropsychology and neuroimaging" project (CA) led by Dra. Andrea Slachevsky and ii) breast cancer information Fondef N. D11I1029 project "Incorporating Next Generation Sequencing Care in cancer Patients" (CC) led by Dra. Katherine Marcelain.

Clinical data collection came from hospital medical records, along with demographic data (CA only). Genomic data was obtained from analysis files Fastq blood samples and / or tissue processed using next-generation DNA sequencing (NGS) (CC).

To adapt us to the dissimilar nature of the recorded data, the information was stored in a new hybrid database system data, allowing both clinical structured data and genomic non structured document type.

The resulting implementation has a filtering system and search terms in bibliographic databases and genomic information in biological databases; Pubmed, RefSeqGene, MedGen, dbSNP, Clinvar, Cosmic, Gene and it is posible to add other resources as needed.

The aim of this thesis is to design and implement a set of software tools to allow extraction, transformation and loading (ETL) of information on databases created and allow online consultations via webservice. These best practices webservice interface design, prototyping and strong development techniques xtreme programming will be built using open source software.

The final goal is that the resulting information is available remotely via a platform that can be accessed from any system using webservice and associated clinical record.

1 Introducción

Durante los últimos años la medicina traslacional o personalizada ha surgido como un enfoque potente para el estudio de enfermedades complejas. Definida por la National Health Institute (NHI) y la Food and Drug Administration de Estados Unidos (FDA) como "una práctica emergente de la medicina que utiliza el perfil genético individual para guiar las decisiones de prevención, diagnóstico, tratamiento de una enfermedad"[1]. La idea fundamental es fortalecer la retroalimentación entre los estudios en ciencias básicas y la clínica para mejorar los diagnósticos y tratamientos de los pacientes. Esta disciplina en parte se retroalimenta de los avances en ciencias ómicas; áreas como genómica, transcriptómica, proteómica y metabolómica[1].

Accediendo a mayor información del paciente, en particular genómica, se busca definir de mejor manera el fenotipo de su enfermedad y con ello decidir el mejor tratamiento. A modo de ejemplo, en el cáncer de mama se ha determinado un tipo HER2 positivo como una forma muy agresiva causada por sobreexpresión de la proteína HER2 en sus células[1]. Este descubrimiento permitió la creación de medicamentos que se dirigen directamente al HER2 disminuyendo el riesgo de ocurrencia.

En cuanto a las ciencias ómicas, la investigación mediante el análisis de microarrays (transcriptómica) en relación al descubrimiento de niveles de expresión génica y hallazgos de mutaciones a lo largo del genoma, han dado lugar desde el año 2007 a más de 70 terapias con genes aprobados por la FDA en pacientes con cáncer, permitiendo dar un fuerte impulso a la comprensión de dicha enfermedad [2]. Posteriormente "el avance de las tecnologías de siguiente generación (NGS) están revolucionando la forma de investigación a escala genómica, y sus efectos están comenzando a trasladarse a la clínica" [3]. Estas NGS son de gran utilidad en la descripción de Polimorfismos Simples de Nucleótidos (SNP) y Variantes Estructurales (SV) y, como los costos de secuenciación siguen descendiendo a una suma actual aproximada de US$1.000, se avisora una mayor utilización de estas tecnologías en el tratamiento clínico.

[1] Ciencias ómicas. (del griego omes, el todo o completo) esta integrada entre otras disciplinas por transcriptómica: estudio de los genes en el mRNA y su transcripción; proteómica: estudia el conjunto de proteínas en un sistema biológico y metabolómica: investigación del sistema metabólico a nivel molecular[25]

El desarrollo del proyecto genoma humano ha permitido además, un avance más que sustancial en esta verdadera revolución médica, sumando un amplio espectro de entidades desarrollando investigación. A modo de ejemplo, el diagnóstico de cáncer basado en análisis molecular es un sector que tiene un rango de crecimiento anual de 18%, incluso durante 2015 el mercado de diagnósticos moleculares se estima crecerá al 14% en base al aumento de la demanda de medicina personalizada[3]. Este crecimiento presenta el desafío de integrar de forma eficiente la información clínica con la originada desde la investigación, y generar métodos para acceder a estos volúmenes que requieren grandes data warehouse y nuevas formas de obtención de información[4].

Las premisas para estas nuevas investigaciones será el acceso a toda la información disponible, la integración de diversas disciplinas que aportan a la investigación y la integración desde sus múltiples procedencias, su almacenamiento y posterior análisis como se representa en la Figura 1.

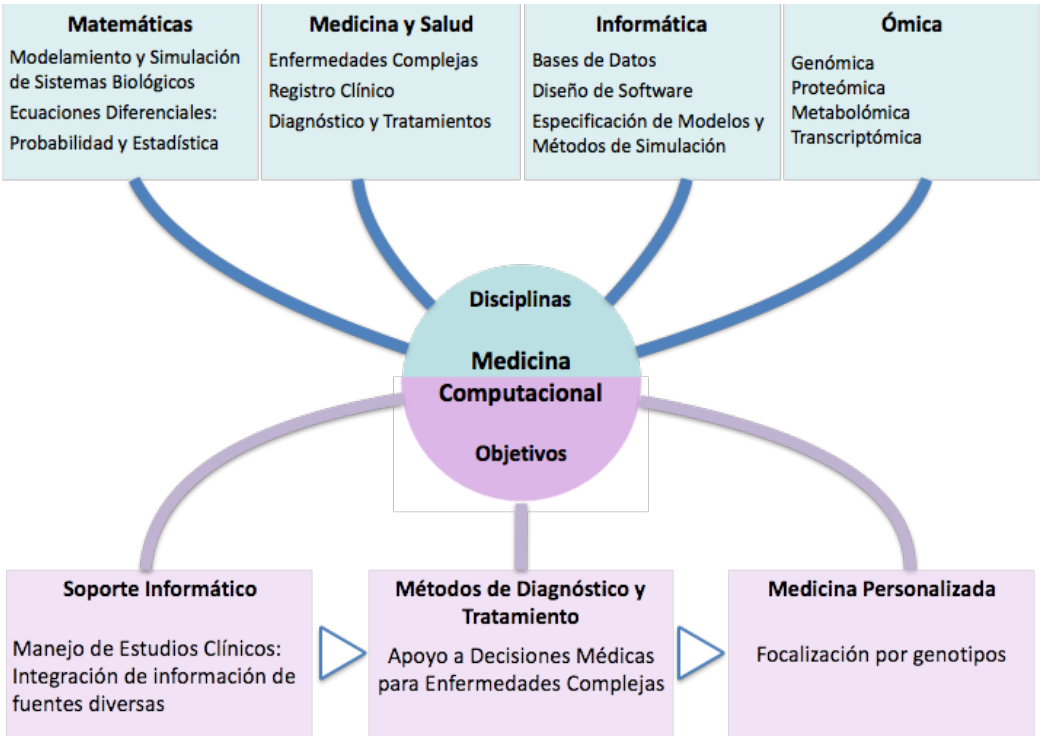

Figura 1. Diagrama de áreas de influencia y metas en medicina computacional.
Matemáticas, Medicina y Salud, Informática, y las ciencias ómicas son disciplinas básicas para la Medicina Computacional, que junto al registro informático apoya diagnósticos y decisiones para lograr medicina personalizada. (Adaptada de la conformación del Centro de Informática Médica y Telemedicina CIMT-U. de Chile).

Dentro de una plataforma de investigación e integración de datos se requiere gestionar o acceder a la mayor heterogeneidad de información posible, y obtener junto a la información genómica la mayor información clínica disponible (diagnósticos, exámenes de laboratorios y signos), generalmente inserta en registros clínicos dentro de las instituciones que otorgan tratamiento al paciente. Esta información se analiza tanto bioinformáticamente (según resultados de secuenciación) como desde la recopilación de antecedentes en bases de datos biomédicas (literatura y bases biológicas), como se muestra en el esquema de la Figura 2.

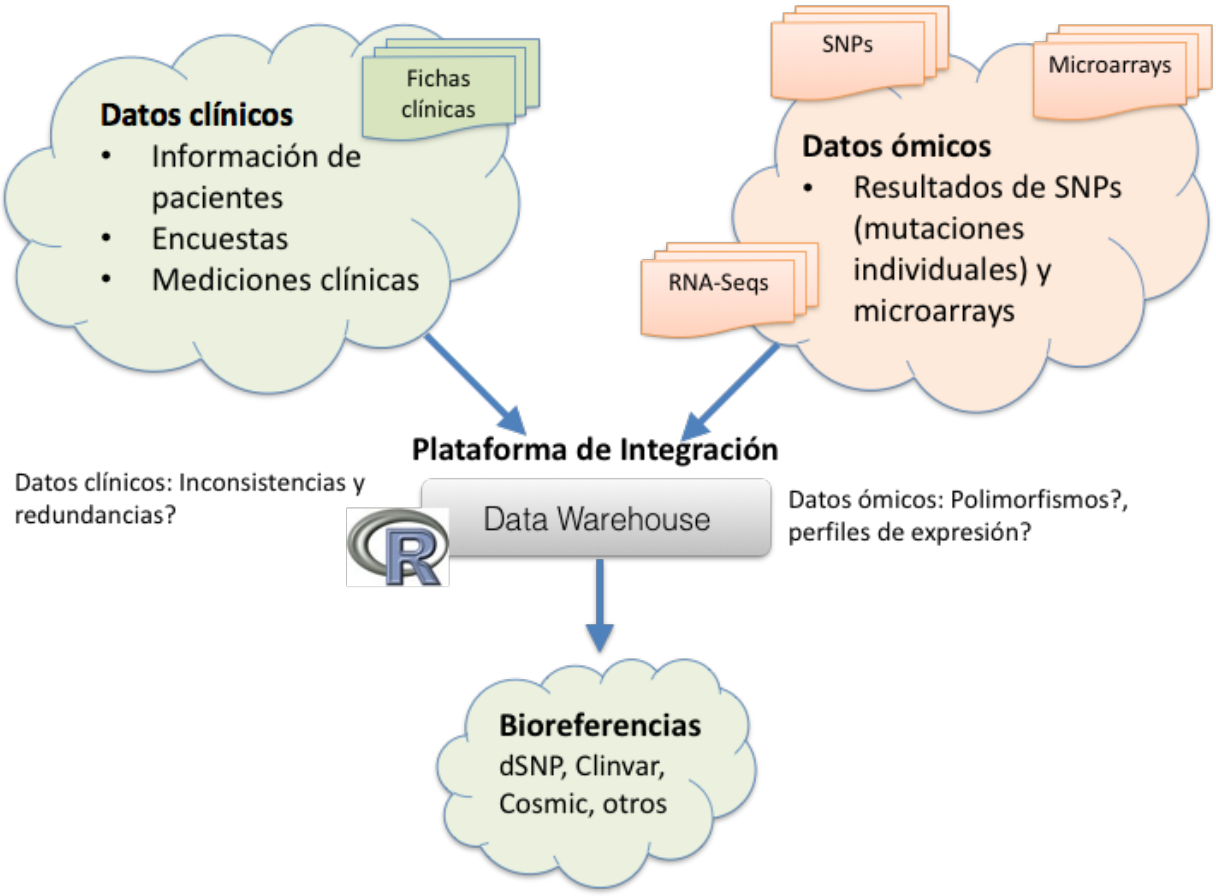

Figura 2. Entidades y flujos para una plataforma de reporte clínico-ómico.
La información tanto clínica por un lado como genómica y transcriptómica, proveniente de SNPs y RNA-Seqs o Microarrays respectivamente, se incorporan en un repositorio (Data Warehouse) para ser analizadas con diversas herramientas estadísticas y/o bioinformáticas y consultadas sobre servicios de referencias externas (Bases de datos biológicas).

1.1 Registros electrónicos

El uso de datos a través de registros electrónicos de salud (EHR por su siglas en inglés) es ampliamente utilizado en el tratamiento y seguimiento clínico de pacientes. Por otra parte, el registro electrónico de descubrimientos genómicos es de uso mas reciente, así como también su aplicación a la investigación de las bases genéticas de las enfermedades y la respuesta al uso de drogas. Los primeros estudios de asociación del genoma (GWAS) se realizaron en 2005 y hacia 2010 se han realizado más de 500 estudios amplios cubriendo cada vez un mayor rango de enfermedades y tratamientos[5].

Los EHR proveen una manera eficiente de analizar diversos fenotipos con sus respectivos genotipos de forma sistemática en números cada vez mayores. A modo de ejemplo, el programa de veteranos de Estados Unidos planea colectar un millón de casos en sus hospitales, El Kaiser Permanent Research Program on Genes busca colectar ADN de 500.000 individuos, al igual que los biobancos de UK y el China Kadoorie biobank[5].

Actualmente los sistema EHR son complejos debido al tipo y volumen de información que administran, de variada índole y de tipo estructurada y semi-estructurada. Al integrar a estos flujos de información la de tipo genómico, la complejidad aumenta considerablemente y se requiere de una serie de adaptaciones, junto con desarrollar nuevos estándares para lograr una buena integración e intercambio de información, tomando en cuenta además que generalmente la información proviene de diversos centros[6].

El advenimiento de todas estas investigaciones en el ámbito médico aunado a los desarrollos en informática médica confluyen hacia una informática de investigación clínica (Clinical research Informatics o CRI)[7], como una sub-disciplina de la informática médica que desarrolla herramientas para reclutamiento, recolección y análisis de datos. Para informática de investigación clínica se requiere aglutinar información de diversa índole, incluyendo: clínica, genómica, ambiental y demográfica con el objetivo de predecir y/o asociar fenotipo y genotipo.

La principal dificultad encontrada al objetivo de aglutinar fuentes de datos para un mejor diagnóstico y/o tratamiento, es la heterogeneidad de estos, y la falta de estándares suficientes para reconocer y almacenar esta información. Existen una serie de estándares en el ámbito de la salud

que pretenden implementar una estandarización, sin embargo, a la hora de decidir cual adoptar no todas las instituciones adoptan la misma.

Los estándares más utilizados en Chile son[1]:

- CIPM: Clasificación Internacional de Procedimientos en Medicina. Basado en CIE-9 modificación Clínica, se utiliza en hospitales del sector público para la generación de Grupos Relacionados de Diagnóstico (GRD). Su objetivo es relacionar el costo de la asistencia medica otorgada.
- CIE-10: Clasificación Internacional de Enfermedades, versión 10. Se utiliza para causas de defunción y determinar la afección principal del egreso hospitalario.
- CIE-O: Clasificación Internacional de Enfermedades, versión oncología. Permite codificar la localización e histología de neoplasias. Esta basado en los informes de anatomía patológica.
- CIF: Clasificación Internacional del funcionamiento. Describe el estado de funcionalidad de la persona atendida y su relación con el estado de salud.
- LOINC: Nombres y Códigos del Identificador para Observación de Laboratorio. Es un conjunto de identificadores y códigos para observaciones clínicas y de laboratorio. Contiene sinónimos, unidades de medida y descripciones detalladas.
- SNOMED-CT: Nomenclatura Sistematizada de Términos de Medicina Clínica. Es una terminología codificada para representar información médica de forma precisa e inequívoca. Se constituye por conceptos, descripciones y relaciones. Es distribuído por la International health terminology Standards Development Organisation (IHTSDO), organización a la que pertenece Chile como miembro[2] desde 2013.

[1] http://www.deis.cl/wp-content/uploads/2011/09/Decreto_Norma_TecnicaEstandares_de_Informacion_DEIS.pdf
[2] http://www.salud-e.cl/servicios-terminologicos/cnomed-ct-chile/.

1.2 Bioinformática y medicina predictiva

El volumen de información genómica tan solo en base a tumores pueden alcanzar más de 10.000 mutaciones y desplazamientos y solo una pequeña porción de estos derivan en células cancerígenas. La potencia computacional es cada vez más requerida y, debido a que los costos tanto de procesamiento como de almacenamiento siguen bajando, es cada vez mas factible implementar plataformas de cómputo de altas prestaciones. Adicionando el incremento del uso de computación en la web, con un mayor nivel de seguridad ha permitido generar estudios a gran escala [8][9]. Junto a lo anterior, se ha desarrollado una amplia variedad de herramientas de software computacional que permiten realizar análisis específicos de todo tipo.

Todo lo anterior ha permitido la creación de una serie de repositorios de datos genéticos, la mayoría de ellos de libre disposición para investigaciones que permiten contrastar resultados tanto de mutaciones como de SNPs en una serie de organismos, como se indica en la Tabla 1, donse se listan algunos repositorios de información.

Tabla 1. Repositorios de informacíon biológica.

Nombre	Estado	Dataset
BioGRID	public	genetic and protein interaction, from organisms and humans. http://thebiogrid.org/
RefSeq	public	Set of sequences; genomic DNA, transcripts and proteins. https://www.ncbi.nlm.nih.gov/refseq/rsg/
Gene	public	Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific. https://www.ncbi.nlm.nih.gov/gene
GeneDB	public	40 genomes (prokaryotes, eukaryotes). http://www.genedb.org
Ensembl	Public	DNA sequences (FASTA) and variants. http://www.ensembl.org

El gran volumen de información es manejado por programas informáticos que son complementados con herramientas informáticas de análisis con un alto grado de exactitud en base a métodos estadísticos de clustering, componentes principales, test de hipótesis (paramétricos y no parámetricos), modelos de markov, machine learning y pattern recognition.

La investigación biológica basada en la información genómica obtenida de la cada vez más extendida secuenciación de ADN ha originado una red compleja de datos de diversos tipos. La complejidad y especificidad de esta información ha generado una serie de bases de consultas, algunas generales, otras muy específicas a la investigación. Las bases generalmente acumulan información de secuenciación, mutaciones y asociaciones con fenotipos de enfermedades de interés.

1.3 Marcos de trabajo integrativo

La gran cantidad y heterogeneidad de los datos que se pueden almacenar en un estudio hace complejo el descubrimiento de información relevante (definir el fenotipo). Para abordar este problema es necesario desarrollar un sistema que permita integrar los estudios realizados a cada paciente y asociar sus resultados. Para lograr esta finalidad se han desarrollado en el tiempo una serie de plataformas que, en alguna medida, abordan ciertos aspectos de la deseada integración de datos para investigación. Dichas plataformas en general deben contemplar una serie de principios fundamentales [10]. Entre ellos se encuentran:

- El respeto a los individuos, promoviendo los derechos e intereses de todos las personas, particularmente de quienes aportan sus datos para investigación biomédica reglamentado en la Ley 20.584[1].
- Promoción de la salud.
- Distribución de beneficios en la forma de cooperación y colaboración internacional en los datos compartidos.
- Reciprocidad, sirviendo como una herramienta de investigación responsable.
- Reproducibilidad, para corroborar las investigaciones asociadas.

Algunas plataformas están orientados en forma específica al estudio de una o un grupo de enfermedades o estudios. A modo de ejemplo, la plataforma BRISK[2] está centrada en el estudio de asociaciones genómicas amplias (GWAS) sobre la base de SNPs y solo soporta datos de dicho tipo. El Portal cBioCancer Genomics contempla un amplio rango de dataset ómicos producidos

[1] http://web.minsal.cl/derechos_deberes_pacientes
[2] http://genapha.icapture.ubc.ca/brisk/index.do

en estudios a gran escala, que incluyen información sobre: mutaciones, alteraciones en copias, microarrays de RNA cambios de expresión en secuenciación de mRNA, metilación de ADN y datos de proteínas y fosfoproteínas.

Los problemas principales se orientan al manejo de la seguridad de los datos clínicos, la capacidad de incorporar datos ómicos, el nivel de soporte en interoperabilidad basados en el soporte de terminologías estándares[11]. Dentro de ellas la más ampliamente utilizada es I2B2 desarrollada por la Informatics for Integrating Biology and Bedside Research Center fundado a su vez por la National Center for Biomedical Computing (NCBC). Actualmente esta plataforma está siendo utilizada por más de cien instituciones y empresas, principalmente en Estados Unidos y en diversos países de Europa y Asia[1].

I2B2 es una plataforma de libre disposición que se basa en una serie de herramientas modulares o celdas que actúan en forma interoperable y que permiten la integración de información sobre un data warehouse[12]–[15]. La potencia de I2B2 se basa en la modularidad de su estructura y la creación de una serie de programas accesorios que pueden interoperar entre ellos.

La implementación de una plataforma tipo I2B2 o TranSmart[2] y su posterior personalización (adaptación de nuevos módulos) es un proceso complejo que toma alrededor de dos a tres años [16], [17] y que excede el alcance de este proyecto.

A partir de la base de I2B2 se crearon otras plataformas similares como SHRINE, que permiten el desarrollo de investigaciones unificando los registros clínicos de pacientes entre múltiples instituciones[18], [19], manteniendo la privacidad de los datos locales de cada institución, y tranSMART, una plataforma abierta para administración de conocimientos e investigación en medicina traslacional e información molecular relativa a ADN, ARN y proteínas.

[1] https://www.i2b2.org/work/i2b2_installations.html
[2] http://transmartfoundation.org

1.4 Modelamiento de la información

1.4.1 Modelo de datos NoSQL y relacional

El mayor reto consiste en aglutinar en forma coherente la vasta información generada, de manera de hacerla eficaz en el tratamiento del paciente. En la fase de adquisición de datos los principales problemas pueden clasificarse en tres categorías: el **volumen** de datos generados originados principalmente de información de análisis genómicos, que pueden llegar a alcanzar con facilidad el nivel de Terabytes (Tb), la **disparidad** de los tipos de información en cuanto a su naturaleza y **normalización** de la información contenida que permita su coherencia estructural a través del tiempo.

El concepto de normalización es bien implementado por las bases de datos relacionales, entre ellas PostgreSQL, que definen estructuras rigídas (esquemas) y logran una alta integridad relacional basada en el teorema CAP (por Consistency, Availability y Partition tolerance)[20] (ver Figura 3). Sin embargo, utilizando este tipo de bases, la variabilidad del tipo de información entre diversos casos de estudio complejiza la mantención y desarrollo de las interfaces de usuario. Las bases de datos relacionales (llamadas tradicionales) presentan dificultad para datos que son altamente cambiantes entre estructuras. Así por ejemplo, una misma "ficha de paciente" para registrar una enfermedad como Cáncer de Mama contiene información de tipo muy distinta a aquella misma "ficha de paciente" que contiene información relevante para un registro de Cáncer de Cólon.

Es en este último punto donde pueden jugar un papel importante las bases sin esquema o NoSQL (Not Only SQL), que son las propuestas para el desarrollo de este proyecto y que son aquellas orientadas a objetos documentos, que no requieren ser definidas previamente, sino que se adaptan según la información almacenada. Esto permite que en la misma tabla "de fichas de paciente", por ejemplo, se pueda almacenar información relevante para una enfermedad y en otro registro de la misma tabla se almecene información respecto de otra enfermedad muy distinta.

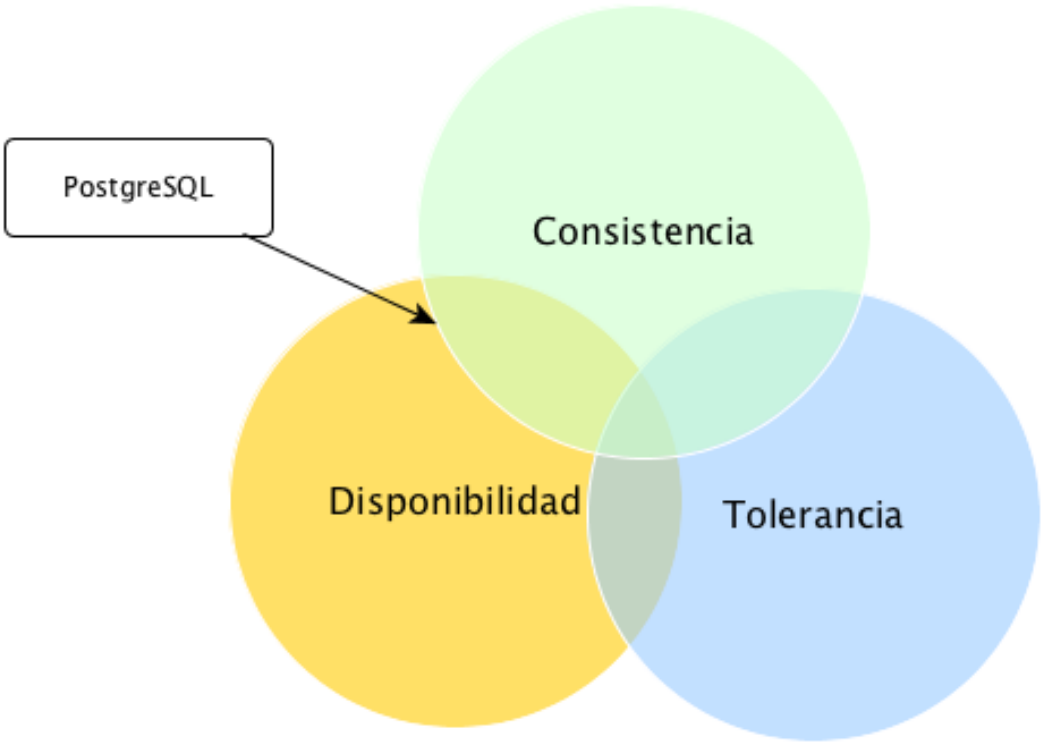

Figura 3. Teorema CAP para bases de datos.
Se indica la posición del motor de base de datos PostgreSQL dentro del teorema CAP. PostgreSQL privilegia la disponibilidad y la consistencia entre tablas mediante claves foráneas, en desmedro de la tolerancia al particionamiento.

Por último, recientemente se ha liberado una versión de una base de datos que podría considerarse "híbrida" (específicamente PostgreSQL versión 9.4 o superior), que puede obtener lo mejor de ambos métodos al flexibilizar la información y permitir a la vez la relación entre tablas mediante claves identificatorias foráneas. Las últimas versiones de este motor de base de datos incluyen soporte para formato de datos json en modo binario (jsonb) y funciones de manipulación de registros para este formato[1].

Como se muestra en la tabla 2, la capacidad del formato json es la flexibilidad en el tipo de información a contener. Se muestra en la columna "NoSQL" cómo dos registros contienen información diferente.

[1] https://www.postgresql.org

Tabla 2. Relación entre las estructuras de una base relacional y NoSQL.
Se muestra la comparación de la flexibilidad de la información en una base relacional y NoSQL.

Relacional	**NoSQL**
Registro 1: “nombre”: “Juan”, ”apellido”:”Perez Soto”, rut:”10363642K”, “edad”:45, “sexo”:”F”, calle”:”San Martin”, “numero”:”13”, ”sector”:””, ”comuna”:”Maipu” **Registro 2:** “nombre”: “Alberto”, ”apellido”:”Araneda Veliz”, rut:”10900642K”, “edad”:48, “sexo”:”F”, calle”:”Anibal Pinto”, “numero”:”1399”, ”sector”:”centro”, ”comuna”:”Maipu”	**Registro 1:** Data:{ “nombre”: “Juan”,”apellido”:”Perez Soto”, rut:”10363642K”, “edad”:45, “sexo”:”F”, dirección:{“calle”:”San Martin”, “numero”:”13”,”sector”:””,”comuna”:”Maipu”} } **Registro 2:** Data:{ “nombre”: “Juan”,”apellido”:”Castro Varas”, rut:”14363642K”, “edad”:48, “sexo”:”F”, dirección:{“calle”:”Orellana”, “numero”:”4590”,”sector”:”Los Vilos”,”comuna”:”Valparaiso”}, “diagnostico”:”xxx”, “glicemia”:”200”, “ingreso”:”2016-10-20” “sangre”:”AB” “calcemia”:”10.7” “APE”:”0.24” }

1.5 Genómica integrativa

La propuesta frente a los puntos anteriormente tratados, conduce a la evaluación de desarrollar una plataforma que permita capturar en forma simple la información, inicialmente de índole exclusivamente genómica, basada en secuenciación de ADN y variantes genéticas junto la información fenotipica de la enfermedad, y junto ello, determinar una forma de almacenar esta información de manera eficiente y que permita la utilización de diversos dominios de información dentro de la misma estructuras de datos.

2 Hipótesis

Se pueden integrar en una plataforma computacional única datos clínicos y genómicos de un paciente y, con ello, establecer relaciones entre mutaciones de determinados genes y el fenotipo de la enfermedad en estudio a través de búsquedas en literatura y análisis estadísticos.

3 Objetivos

3.1 Objetivo General

Crear un modelo de plataforma de integración de datos clínicos y genómicos para estudios en Alzheimer y cáncer de mama.

Se implementará una plataforma (Datagenomed) que permita integrar información de resultados genómicos de SNPs con información clínica. Para ello, se diseñará e implementará una interface de captura de datos para pacientes con Alzheimer y cáncer de mama, incorporando las herramientas necesarias para análisis estadístico con R.

Las consultas de información se implementaran en un modelo que permita integrar las bases de datos clínicas y las bases de datos genómicos.

3.2 Objetivos Específicos

1) Implementar la utilización de un repositorio de datos (data warehouse) para datos clínicos y genómicos.
    a) Seleccionar arquitecturas de bases de datos clínicas y genómicas.
    b) Definir la estructura de almacenamiento estandarizado de datos clínicos en Alzheimer y cáncer: definir elementos comunes y específicos.

c) Determinar las estructura de datos para información genómica en base a archivos de secuencia Fastq.

d) Almacenar datos procedentes de secuenciaciones géneticas.

e) Construir un modelo de sistema de almacenamiento integrado de datos clínicos y genómicos.

2) Diseñar e implementar un sistema software integrado clínico-genómico.

a) Diseñar un sistema software de integración de información clínica y genómica.

b) Construir un sistema software de integración de información clínica y genómica.

3) Construir interfaces hacia plataformas de integración de información.

a) Especificar requisitos de desarrollo de herramientas ETL (extracción, transformación y carga) sobre la base de datos clinico-genómicos.

b) Usar una interface de extracción, transformación y carga de datos (ETL) hacia repositorios especializados.

c) Construir interfaces hacia R para análisis estadístico.

d) Evaluar el software a través de encuestas de usabilidad aplicadas a casos de estudio.

Materiales y métodos

4 Materiales y Métodos

Establecido el propósito de construir y evaluar una plataforma de software para el tipo de información planteado, se definieron una serie de etapas entre las que se mencionan la definición de casos a utilizar, analizar el tipo de información capturada de cada caso para definir la estructura de datos apropiada y finalmente construir el software requerido. Se describen las etapas de la construcción de la plataforma siguiendo las pautas de construcción de software: definiendo requerimientos, características de diseño conceptual, para finalmente presentar las interfaces logradas en el software. Cada etapa fue posible gracias a una activa comunicación con los usuarios directos aquí considerados. En particular, de dicha comunicación se obtuvo el diagrama mostrado en la Figura 4, describiendo las áreas involucrados y sobre todo la información relevante a cada una de ellas.

4.1 Definición de casos

El inicio del proceso de extracción de datos clínicos comenzó con la definición de un conjunto de datos comunes a los casos a evaluar. El conjunto de datos establecidos, agrupados en las categorías demográficos, clínicos generales y especiales, siendo estos últimos referentes a las patologías analizadas.

Definida la información relevante de cada grupo de pacientes y/o muestras, se estableció un mapeo de los conceptos locales de origen con la estructura definida para esta plataforma. En el caso de incongruencias se mapeó la información en base a la nomenclatura CIE-10.

En esta actividad se desarrolló una plataforma (Datagenomed) constituida por un modelo de base de datos y un conjunto de herramientas computacionales que permiten asociar la información clínica del paciente con la información genómica.

La plataforma se adaptó a dos casos de estudio: Alzheimer (CA) y cáncer de mama (CC).

La información referente a pacientes con cáncer de mama proceden del proyecto Fondef No. D11I1029 “Incorporación de la Secuenciación de Última Generación en el Cuidado de los

Pacientes con Cáncer" proporcionado por la Dra. Katherine Marcelain. Dicho proyecto cuenta con cerca de 190 pacientes con información de datos clínicos y genómicos.

La información de pacientes con Alzheimer se obtuvo a partir de una serie de formularios diseñados para este fin y aplicados a una cohorte de pacientes en estudio dentro de un proyecto Fondecyt No. 1140423 a cargo de la Dra. Andrea Slachesvky. Para estos pacientes se cuenta con el conocimiento informado.

Los datos clínicos se obtuvieron mediante recolección por fichas clínicas hospitalarias que, junto a datos demográficos, incluye mediciones clínicas. Los datos genómicos se obtuvieron del análisis de archivos Fastq de muestras de tejido (sangre/tejido) procesados mediante next-generation DNA sequencing (NGS) (CC).

Para adaptarse a la naturaleza heterogénea de los datos consultados, la información se almacenó en un sistema de bases de datos híbrido, permitiendo tanto datos clínicos estructurados como datos genómicos de tipo documental (que sin esquema previo dan mayor flexibilidad).

La arquitectura propuesta se incorporó dentro del flujo de trabajo actual del Programa de Genética Humana, como se indica en la Figura 5, donde la disposición de Datagenomed se situó entre la captura de datos genómicos y la entrega de resultados al(los) investigador(es).

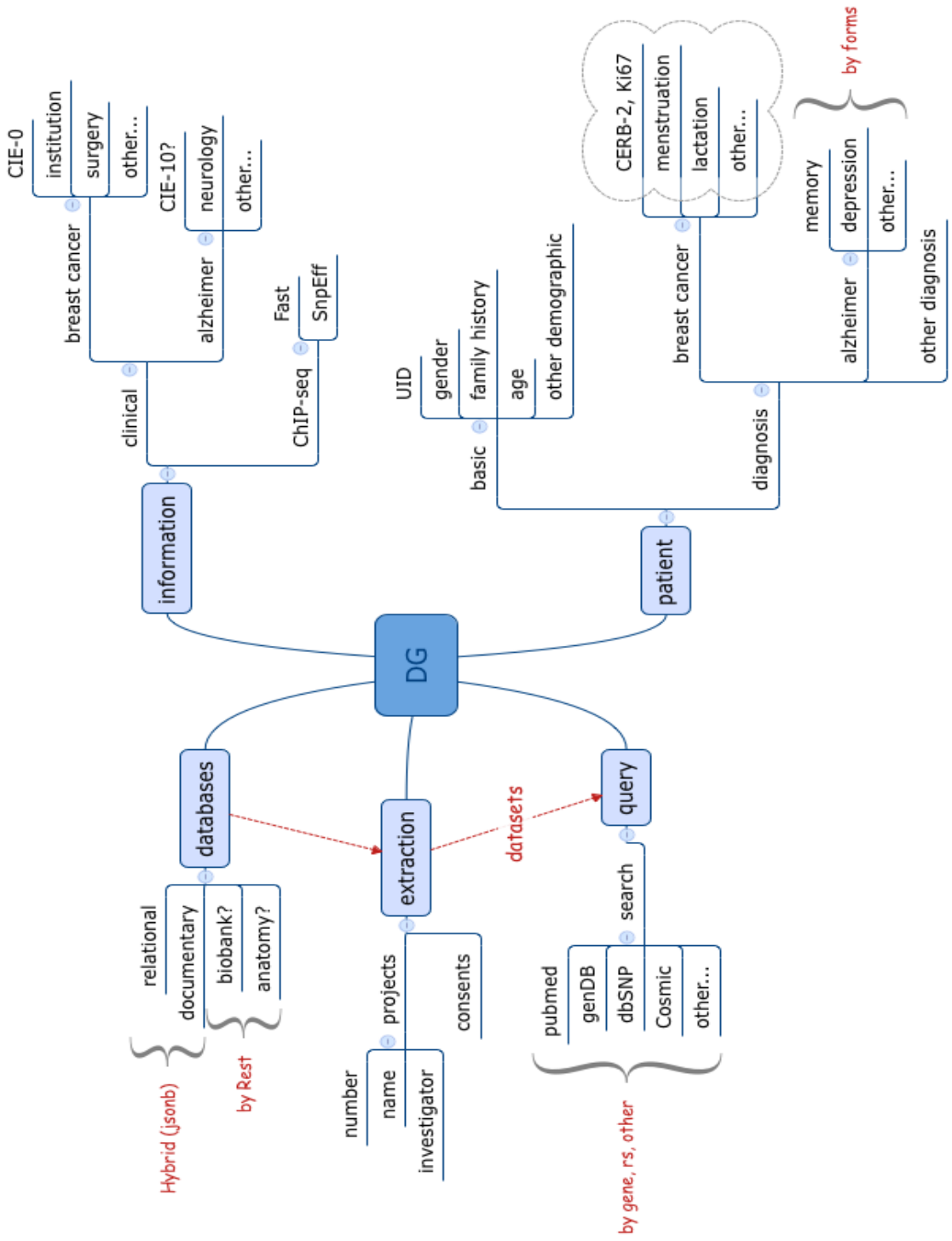

Figura 4. Diagrama de información en plataforma Datagenomed.

Se indica el tipo de información a capturar/almacenar en cada modulo dentro de Datagenomed.

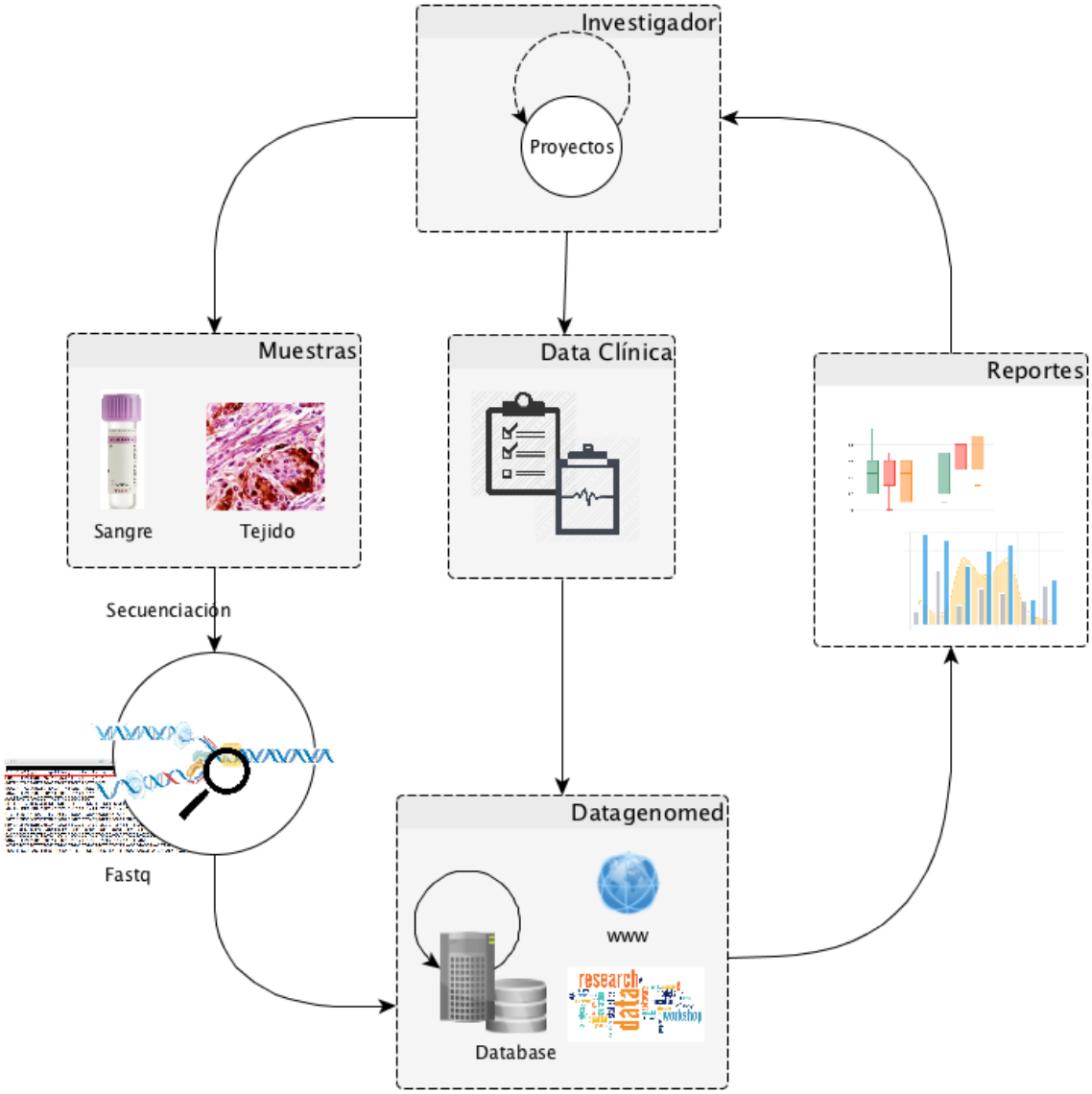

Figura 5. Relación de elementos en plataforma genómica.
La información colectada, tanto clínica como genómica se instala en Datagenomed para ser procesada y enviada a los investigadores en forma de reportes y consultas a bases biológicas referenciales.

Una vez descubiertas asociaciones clínico-genómicas se interactuó con bases de datos genómicas externas ya sea para corroborar dichas relaciones o para inferir nuevas formas de diagnóstico y tratamiento[17].

Se diseñó e implementó un conjunto de herramientas de software para permitir procesos de consulta de información sobre las bases de datos creadas y permitir accesos en línea a usuarios autentificados.

Se evaluaron tres plataformas open source para implementar procesos de extracción, transformación y carga de datos (ETL), entre ellos: Pentahoo, Talend y Knime, de ellas se seleccionó Knime[1] por contener una serie de modulos de conexión a R en forma nativa y a bases de datos PostgreSQL y modulos especialmente diseñados para su uso en bioinformática.

Knime también ofrece una amplia cantidad de herramientas estadísticas (test de hipótesis, correlación linear, t-test, anova, predictor de regresión, entre otras), y de minería de datos (clustering, arboles de decisión, análisis de componentes principales, etc.). En la Figura 6 se ejemplifica un flujo de trabajo simple para depurar un archivo Excel.

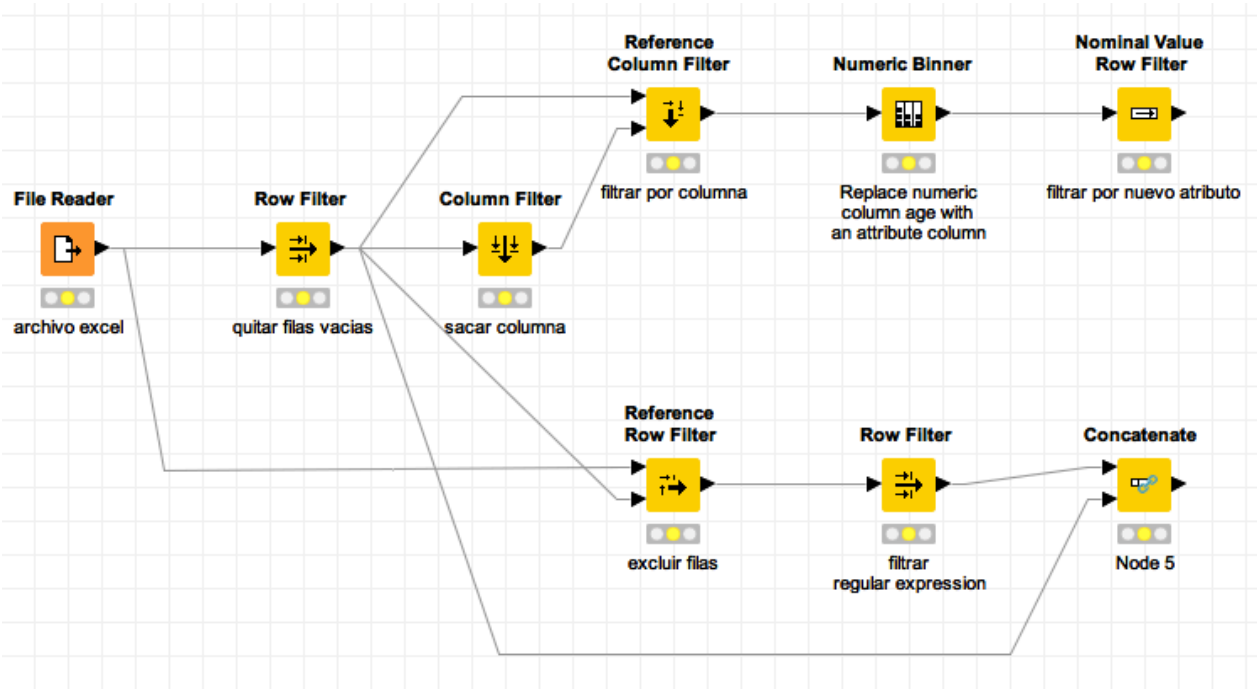

Figura 6. Ejemplo de flujo de trabajo en Knime.
Secuencia de procesos para cargar un archivo en formato Excel. Siguiendo la ruta inferior del fllujo; a la carga inicial del archivo se aplican una serie de procesos para eliminar filas vacías y luego eliminar filas por un criterio definido, posteriormente se aplica otro filtro utilizando expresiones regulares para seleccionar filas, y finalmente, obtener un archivo resultante depurado.

[1] http://www.knime.org

4.2 Evaluación de software

Para el cumplimiento del objetivo 3.d se realizaron modificaciones para implementar metodologías de evaluación de usabilidad, basadas en la Norma ISO /IEC 9126-4[1]- de usabilidad métrica y en indicaciones de usabilidad en entornos web y mediciones de usabilidad[21]-.

Se generó un cuestionario de evaluación de usabilidad que se aplicó a los usuarios respectivos y luego se analizó estadísticamente. Con ello se detectaron las fortalezas y debilidades de la plataforma Datagenomed. Mayor detalle de las encuestas en Anexo 1.

4.3 Especificación de requisitos

4.3.1 Usuarios del software: Usuarios directos

Se identificaron como usuarios directos del sistema a aquellos que, pertenecientes a la institución, utilizan el sistema para el ingreso/edición de información y que acceden desde dentro de la institución central y/o desde Internet. En este contexto son usuarios:

Secretaria/enfermera:

Personal encargado de la recepción de los datos de casos y quién debe registrar la información detallada de éstas. Digita la información determinada por el supervisor en los registros clínicos. El acceso de estos usuarios es a través de la carga inicial de datos desde archivos de textos (exportados a su vez desde documentos Microsoft Excel) o desde la ficha de ingreso de casos. También es quien ingresa y registra el consentimiento informado si corresponde.

Supervisor:

Profesional encargado de la revisión de las solicitudes de casos. Accede a la ficha de diagnóstico de casos.

[1] El estándar ISO/IEC 9216-4 se centra en métricas para la evaluación de calidad en ingeniería de software.

Laborante/técnico:

Profesional de laboratorio de genética que procesa mediante secuenciación, las muestras asociadas a cada caso para la obtención de archivos Fastq. Revisa y corrobora la información emitida por Secretaria.

Bioinformático:

Profesional especializado que procesa y analiza las muestras obtenidas por secuenciación y genera archivos resultantes en formato VCF que carga al sistema.

Investigador responsable:

Este usuario realiza las consultas y extracciones de información para consultas bibliográficas y de asociación genómica. Evalúa y filtra las secuencias según necesidad y accede a las bases de datos científicas predefinidas en el sistema.

4.3.2 Usuarios del software: Usuarios indirectos

Los usuarios indirectos del sistema son:

- personal médico asociado a las instituciones de salud,
- participantes de cada proyecto

Ambos tipos de usuarios consultan los informes de pacientes desde la página web del sistema, o recepcionan en sus correos electrónicos los informes elaborados por el investigador.

Los usuarios señalados anteriormente accederán al sistema utilizando su rut como nombre de usuario y una clave que le será asignada.

4.3.3 Modelos de casos de uso

Las muestras originadas en la clínica u hospital fueron enviadas al Laboratorio de Patología Molecular y Genómica del Cáncer para su secuenciación mediante procedimiento NGS, junto con

ello se enviaron y/o registraron los datos del caso y antecedentes del paciente. Esa información fue ingresada por secretaria y revisada por el supervisor del laboratorio de secuenciación.

El personal técnico de laboratorio realizó el proceso de secuenciación y se enviaron los resultados a análisis bioinformático donde se realizaron los procesos de depuración. El personal bioinformático realizó la carga en un archivo de texto con la información en formato VCF en el sistema Datagenomed asociando el archivo con el caso de estudio. El proceso se esquematiza en la Figura 7.

La información cargada en el sistema es dispuesta para consulta por el investigador y/o el resto de usuarios del sistema.

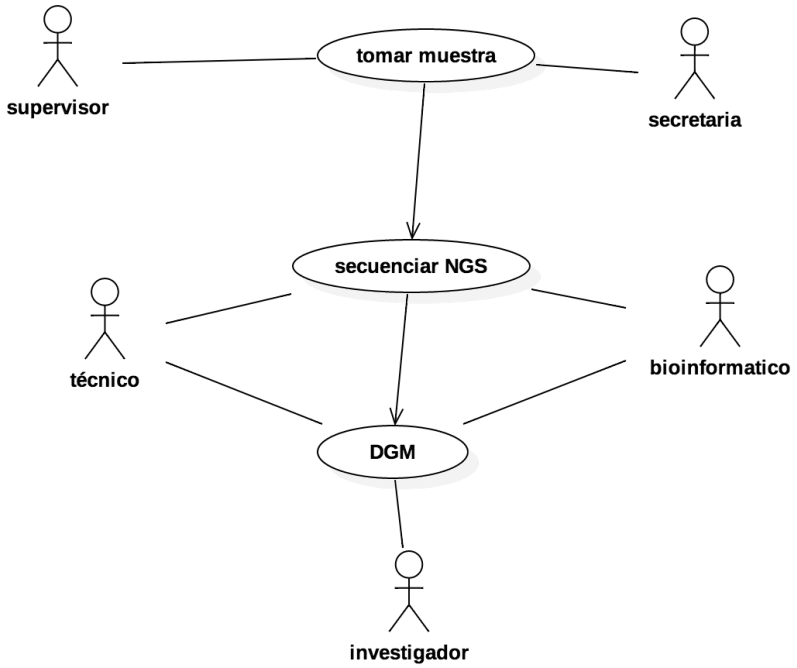

Figura 7. Caso de uso general.
Las muestras obtenidas son posteriormente secuenciadas mediante NGS y sus resultados ya procesados son depositados en DGM para ser consultados.

Las funciones de búsqueda se asociaron a repositorios preestablecidos, con la opción de agregar una mayor cantidad de dichos repositorios en la medida que sea necesario. En la Figura 8 se detalla la integración de DGM en este proceso donde se indica algunas de las referencias predefinidas.

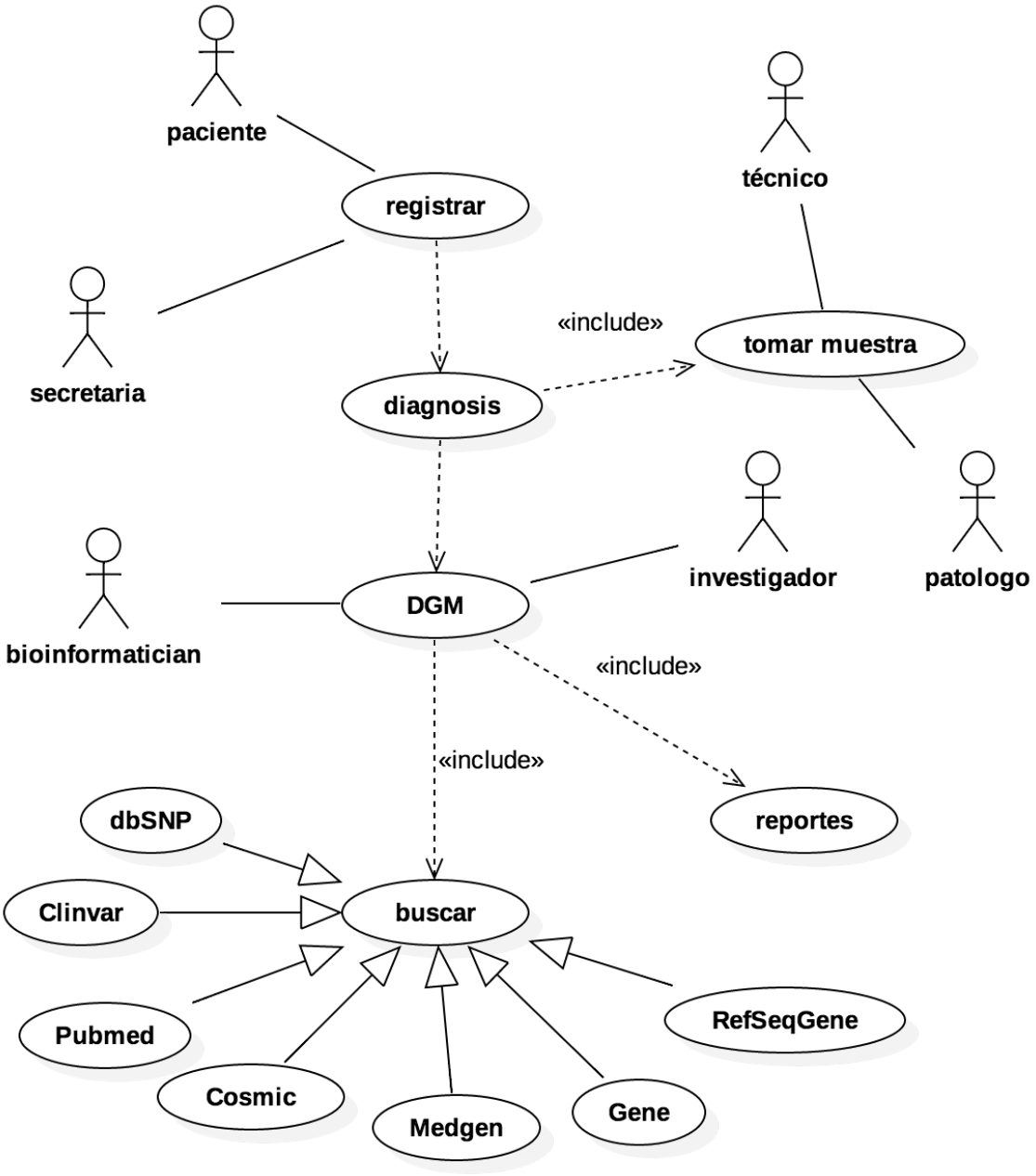

Figura 8. Caso de uso general conexión a referencias.

Se inicia el proceso con el registro de pacientes y toma de muestras de aquellos que cumplen ciertos criterios de diagnosis. Datagenomed (DGM) recopila la información resultante, desde donde se realizan las búsquedas hacia bases públicas de información biomédica (bases de datos biológicas) y se generan reportes de los resultados obtenidos.

4.3.4 Requisitos funcionales

Como resultado de reuniones periódicas con los usuarios se establecieron los siguientes elementos como requisitos del software a implementar, considerando las áreas de servicio a utilizar. Cada categoría de información contiene su mantenedor de información, definiéndose los siguientes:

Mantenedores

- M01. Administrar usuarios
- M02. Administrar pacientes
- M03. Administrar proyectos
- M04. Administrar secuencias
- M05. Administrar casos
- M06. Administrar grupos
- M07. Administrar consultas

Consultas

- C01. Consulta dataset de casos
- C02. Consulta secuencias
- C03. Consulta SnpEff
- C04. Consulta bibliografía
- C05. Consulta dbSNP
- C06. Consulta Cosmic
- C07. Consulta Clinvar
- C08. Consulta RefSeqGene

4.3.5 Mantenedores

Para todos los mantenedores definidos en el sistema se validó que los datos indicados como obligatorios sean efectivamente ingresados, emitiendo un mensaje de alerta al usuario en caso contrario.

La información relacionada con el rut o identificador en cada mantenedor que lo utilice debe ser validada y comprobar su existencia dentro del sistema en caso requerido.

M01. Administrar usuarios

El sistema maneja un formulario de ingresos de usuarios, donde el usuario administrador ingresa y/o modifica los atributos de éstos.

- Existe un formulario específico para el ingreso de usuarios.
- Los datos que se deben ingresar son: rut, nombre, apellido, teléfono, celular, mail, rol, grupo primario de pertenencia, grupos secundarios de pertenencia, estado.
- Los datos obligatorios son el rut, nombre y apellido, email, estado.
- Si falta algún campo por ingresar y que no sea opcional, se ser emitido un mensaje de alerta para el usuario.
- Se valida que al ingresar un nuevo rut, éste no exista previamente.
- Se valida que el rut ingresado sea válido.

M02. Administrar pacientes

Formulario para administrar la información relativa a los pacientes (sujetos) que poseen casos de estudio dentro del sistema.

- Los datos a ingresar son iniciales, edad, sexo.
- La información obligatoria son iniciales, edad, sexo.

M03. Administrar proyectos

Formulario destinado a registrar los proyectos de investigación aprobados para uso de muestras.

- Los datos a ingresar son nombre del proyecto, código fondecyt, nombre y apellido del investigador principal (PI), institución origen del proyecto.
- Información adicional: objetivos del proyecto, tipo de financiamiento; público, privado o mixto.
- Todos los datos son obligatorios.

M04. Administrar secuencias

Formulario para administrar la información relativa a la secuencia obtenida en el proceso. Se carga en forma VCF.

- Carga de archivo en formato VCF en base de datos.
- La modificación de la información colectada no esta habilitada.

M05. Administrar casos

Formulario para administrar la información de los casos (sujetos) dentro del sistema.

- Los datos a ingresar son iniciales, edad, sexo.
- La información obligatoria son id de consentimiento, origen, diagnóstico,

La información requerida varía según sea el tipo de caso, debido al diagnóstico diferenciado e información requerida tanto en caso de Alzheimer y/o Cáncer de mama.

Los consentimientos deben estar asociados al proyecto que solicita la información, por ende existen diversos tipos de consentimientos:

- universal: están abiertos a ser usados en cualquier proyecto de investigación.
- exclusiva: están disponibles única y exclusivamente al proyecto que fue concedido.

M06. Administrar grupos

Formulario para gestionar grupos de usuarios y/o instituciones de salud participantes de proyectos de investigación.

Información a ingresar: id, nombre,

M07. Administrar consultas

Gestión de consultas y asignación de estas a grupos de usuarios.

- Información a ingresar: id, datos del filtrado de casos, fecha, nombre de la consulta, id de grupo, id de usuario.
- Información obligatoria: fecha, nombre de la consulta, id de grupo, id de usuario.

4.3.6 Consultas

Todas las consultas se definieron ser ejecutadas directamente sobre el/los formularios disponibles en pantalla.

C01. Consulta casos

Para la consulta de casos se dispone de una serie de filtros de búsqueda que permitirá buscar por varios campos de información (edad, sexo, diagnostico, detalle, RE, RP, CERB2, Ki67). La información demográfica estará limitada a lo existente dentro de cada caso. En CC solo se dispone de la edad y sexo del paciente como información demográfica, en cambio en CA estará disponible una información más completa, incluyendo rut y dirección.

Se emite un listado mostrando data completa de la información referida a continuación según el tipo de caso:

CC (cáncer):

origen, sujeto, edad, peso, Aco, menopausia, diagnóstico, detalle, RE, RP, CERB2, KI67.

CA (Alzheimer):

origen, sujeto, edad,

C02. Consulta secuencias

La consulta de secuencias esta supeditada a la selección previa de casos en consulta descrita en C01. Una vez seleccionada el subconjunto de datos, es factible consultar el total de secuencias

para el subset de casos o para un caso particular. La información a visualizar es la totalidad dispuesta en la base de datos.

El filtrado de secuencias permite seleccionar uno o más cromosomas, uno o más genes, y hasta tres claves adicionales. Todo el filtrado se aplica sobre la selección anterior de casos.

C03. Consulta SnpEff

Consulta de resultados de variantes genéticas. Despliega el resultado de la aplicación de la herramienta SnpEff[22] sobre la secuenciación, para anotar variantes mostrando entre otra información: alelo, efecto de gen, biotipo de transcripción, etc.

C04. Consulta bibliografías

Desde la consulta de secuencias (C02) se despliega las consultas de antecedentes bibliográficos en Pubmed, Gene y MedGen, desde la selección del gen respectivo en la lista de secuencias.

Se activan desde la lista consulta de secuencias el id de gen (ej: rs1873778) y nombre gen (ej: FGFR3) para ser consultados en referencia bibliográfica y/o repositorio de datos.

C05. Consulta dbSNP

Consulta de secuencias por variantes del tipo rs17846816 en base de datos dbSNP u otras a predeterminar.

C06. Consulta Cosmic

Desde la consulta de secuencias (C02) se desplegará las consultas sobre Cosmic, desde la selección del gen respectivo en la lista de secuencias.

C07. Consulta Clinvar

Desde la consulta de snpeff (C03) se desplegará las consultas sobre Clinvar, desde la selección de la ubicación de gen respectivo en la lista.

C08. Consulta RefSeqGene

Consulta de secuencias por genes en base a nombre, código o numeración rs.

4.3.7 Requisitos no funcionales

El sistema se implementa actualmente sobre equipamiento de la Facultad de Medicina y puede ser visualizado en cualquier navegador web.

Se requiere de una conexión a internet para el uso del sistema. Se implementó sobre una base de datos PostgreSQL versión 9.4 con un uso de formato de datos jsonb. Implementado sobre plataforma Linux en una distribución Ubuntu 14.04 LTS.

Facilidad de uso

El sistema en cuestión debe presentar una interface simplificada para el usuario. Basado en las prácticas de usabilidad utilizando la librería Bootstrap versión 3.0 para implementar una interface responsiva entre diversos dispositivos de escritorio y dispositivos móviles (tablets) excluyendo dispositivos celulares.

Portabilidad

Se requiere un navegador web instalado tanto en un dispositivo móvil como en un PC de escritorio. Puede ser utilizado desde cualquier navegador, ya sea Firefox, Chrome, Internet Explorer, Safari, operando bajo los sistemas operativos Windows, Mac OS X o Linux.

En una primera etapa se trabajará preferentemente bajo navegador Firefox, para posteriormente evaluar las diferencias de funcionamiento con otros navegadores. Se evitará en todo caso el uso de Internet Explorer debido a su falta de soporte a los estándares utilizados por otros navegadores.

Respaldos

Se establecen respaldos automáticos e incrementales que son programados en forma diaria, semanal y mensual. Estos son generados utilizando script Bash en el servidor y los archivos resultantes comprimidos en formato Gunzip. Los archivos serán enviados a su vez a discos de respaldo externos asociados al servidor.

Ante un fallo en el software del sistema, existe un período de un máximo de 45 minutos (realizado por un operador previamente capacitado) para restaurar los datos y programas y poner en marcha el sistema.

Impresión

Todos los resultados de reportes serán impresos a una resolución mínima de 300dpi para obtener una calidad de impresión en impresora tipo láser. El tiempo de respuesta o latencia no será mayor a 20 segundos cuando haya hasta 5 usuarios simultáneos accediendo al sistema.

4.3.8 Requisitos técnicos

Dada la naturaleza de la solución planteada, se define el funcionamiento en esta fase, en una red interna dentro de las dependencias de la institución y que el servidor del sistema tenga asignada una IP estática dentro de la red local. Para la conexión con usuarios externos, es necesario que exista una red LAN estable dentro de la institución y con conexión a internet para el acceso de usuarios externos.

Para establecer conexión con usuarios externos vía internet se implementa una DNS de forma de establecer un dominio virtual del tipo www.datagenomed.cl u otro disponible.

Tanto las aplicaciones de software como las bases de datos se implementan en forma inicial sobre un servidor Hewlett-Packard Proliant DL120 operando con sistema operativo Ubuntu Linux 14.04

Especificaciones de software

Los programas utilizados para la creación de esta plataform son:

- Sistema Operativo: Ubuntu Linux 14.04 LTS
- Base de datos: PostgreSQL 9.4
- Servidor web: Apache 2
- Librerías javascript: Bootstrap 3, jquery 1.11,
- Lenguaje de programación PHP v5.6
- R Server versión 3.2
- Shiny server versión 1.4.7

4.4 Diseño Conceptual de Datos

Desde el punto de vista funcional para la plataforma propuesta, la estructura de datos se basa en un modelo documental sobre base de datos PostgresSQL version 9.4 en una forma forma **híbrida**[1], utilizando estructura de datos NoSQL.

Se definió que el formato de los datos sea jsonb (json binario) y solo en las tablas necesarias se estableció un identificador de clave primaria (id) con tipo de formato de datos **uuid** utilizando como función generadora la versión 4[2].

La estructura de la base de datos se muestra en el diagrama de entidad relación de la Figura 9, donde se detalla solo a modo de referencia los campos contenidos en alguna de ellas, ya que la estructura de datos jsonb no requiere de este detalle de esquema en las tablas.

[1] Se utiliza un formato de bases de datos relacional tradicional en forma simultánea con un formato no esquematizado (NoSQL) sobre la misma plataforma de base de datos.
[2] https://www.ietf.org/rfc/rfc4122.txt

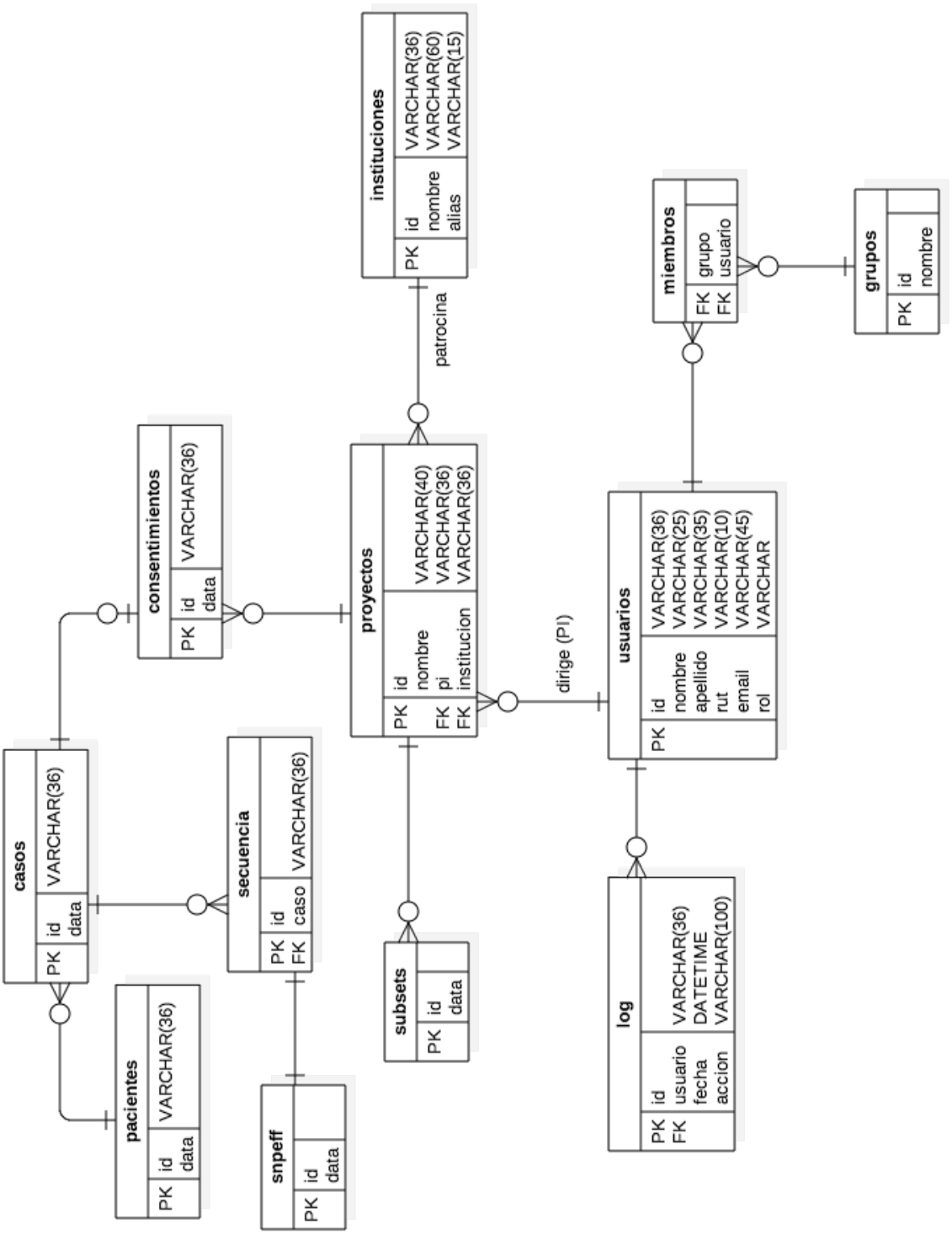

Figura 9. Modelo de entidad relación de la plataforma Datagenomed.

Se muestra en algunos casos a modo de ejemplo, la información de campos contenidos en las tablas (usuarios, proyectos), aunque la totalidad de las tablas utilizan un campo “data” con formato jsonb.

4.4.1 Documentación del Diseño Conceptual de Datos

Usuarios

Atributos

Nombre	usuarios		
Descripción	usuarios del sistema		
Nombre	Tipo de Datos	Descripción	Dominio
data	jsonb	Data completa de usuario	Array json válido
id	uuid	Identificador universal	string

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
id	log	usuarios	1	1..*

Restricciones:

- Cada usuario debe tener un nombre, rut, password y correo como mínimo.
- Todo correo debe constar con un nombre y un dominio separados por un carácter “@”
- El rut debe ser válido y contar con dígito verificador.
- EL rut se almacena en formato de texto con dígito verificador incluído.

Ejemplo data:

```
{
"gid":”60219288-9dda-4ebc-a4e9-f134d3a87536”,
“rut”:”9530872K”,
"email": "rodrigo.vega@gmail.com",
"nombre": "Juan”,
“apellido”:”Soto",
"disabled": 0,
"password": “9a51be8dc6d3003f6f0a4df7c93e017a"
}
```

Grupos

Atributos

Nombre	grupos		
Descripción	grupos de usuarios del sistema		
Nombre	Tipo de Datos	Descripción	Dominio
data	jsonb	Data completa de usuario	Array json válido
id	uuid	Identificador universal	string

Asociaciones

1.1 Nombre	1.2 Clase B	1.3 Clase A	Cardinalidad A	Cardinalidad B
1.6 id	1.7 log	1.8 usuarios	1.9 1	1.10 1..*

Restricciones:

Cada grupo debe tener un nombre, id como mínimo.

Ejemplo data:

```
{
"gid":”60219288-9dda-4ebc-a4e9-f134d3a87536”,
"nombre": "Universidad Andres Bello”,
“alias”:”UNAB",
"disabled": 0
}
```

Miembros

Atributos

Nombre	miembros		
Descripción	usuarios y grupos asociados		
Nombre	Tipo de Datos	Descripción	Dominio
grupo	uuid	Identificador universal	string
usuario	uuid	Identificador universal	string

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
grupo	miembros	grupos	1	1..*
usuario	miembros	usuarios	1	1..*

Restricciones:

- Cada registro debe tener un id, usuario y grupo como mínimo.

Ejemplo data:

```
{
"gid":”60219288-9dda-4ebc-a4e9-f134d3a87536”,
“usuario”:” 60219288-9dda-4ebc-a4e9-f134d3a87536"
}
```

Instituciones

Atributos

Nombre	instituciones		
Descripción	Instituciones que registran proyectos de investigación		
Nombre	Tipo de Datos	Descripción	Dominio
data	jsonb	Data completa de usuario	Array json válido
id	uuid	Identificador universal	string

Asociaciones

Nombre	Clase A	Clase B	Cardinalidad A	Cardinalidad B
id	instituciones	Proyectos	1	0..*

Restricciones:

- Cada institución debe tener un nombre, rut.
- Información adicional: dirección, nombre de contacto, telefono de contacto, alias.
- Todo correo debe constar con un nombre y un dominio separados por un carácter “@”
- El rut debe ser válido y contar con dígito verificador.
- EL rut se almacena en formato de texto con dígito verificador incluído.

Ejemplo data:

```
{"id":”60219288-9dda-4ebc-a4e9-f134d3a87536”, “rut”:”9530872K”,
"email": "rodrigo.vega@gmail.com", "nombre": "Universidad Andres Bello”,
“alias”:”UNAB",
"disabled": 0,
"contacto": “Juan Valdes",
“direccion”:”Anibal Pinto 300”,
“ciudad”:”Santiago”}
```

Proyectos

Atributos

Nombre	proyectos		
Descripción	Proyecto de investigación autorizado		
Nombre	Tipo de Datos	Descripción	Dominio
id	uuid	El identificador único	Valor string único positivo
data	jsonb	Data de proyecto	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
idp	consentimientos	proyectos	1	0..*

Restricciones:

- El nombre y código fondecyt es requisito.
- Cada proyecto debe tener un nombre de investigador responsable.
- El nombre del investigador se indica completo en la forma; “apellido”, “nombres” .
- El teléfono de cualquier clase debe tener número de área y número de teléfono en el formato 99-99999999.
- El correo (email) debe constar con un nombre y un dominio separados por un carácter “@”

Ejemplo data:

```
{
“codigo”:”CO-12390”,
“nombre”:”Depresion y Funcionalidad en pacientes con Alzheimer”,
“PI”:”Navarro, Alejandro”,
“prefijo”:”Dr.”,
“telefono”:”+56978872846”,
“email”:”proyecto@med.uchile.cl”
}
```

Log

Atributos

Nombre	Log		
Descripción	Registro de actividad en el sistema		
Nombre	Tipo de Datos	Descripción	Dominio
Data	jsonb	Data completa de registro de log	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
usuario	usuarios	log	1..*	1

Restricciones:

- Es requisito la fecha, el id del usuario y la acción a registrar.

Ejemplo data:

```
{
"usuario": "60219288-9dda-4ebc-a4e9-f134d3a87536",
"fecha": “2016-03-13”,
"actividad": "update",
"db": " usuarios",
"accion": “delete from usuarios where …”
}
```

Sujetos

Atributos

Nombre	sujetos		
Descripción	pacientes con estudios dentro del sistema		
Nombre	Tipo de Datos	Descripción	Dominio
id	uuid	Identificador serial	Valor string único positivo
data	jsonb	Data de paciente	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
id	pacientes	secuencia	1..*	1

Restricciones:

- Cada paciente debe tener un identificador único de sujeto.
- Debe contener edad y sexo.
- Debe incluir iniciales del nombre.
- Si incorpora rut este debe ser validado y verificado.

Ejemplo data:

```
{"idp": 5,
"edad": 49,
"peso": "62",
"sujeto": 5,
"iniciales": "MLZ"}
```

Casos

Atributos

Nombre	casos		
Descripción	Registro de casos de estudio		
Nombre	Tipo de Datos	Descripción	Dominio
id	uuid	Identificador serial	Valor string único positivo
Data	jsonb	data	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
pid	sujetos	casos	1..*	1
id	secuencias	casos	1	1..*

Restricciones:

- El id de paciente no debe ser nulo.
- Es requisito la fecha de ingreso del caso.
- Debe identificar la institución y el médico tratante.

Ejemplo data:

```
{"ci": "Si",
"re": "Positivo ++ a +++en el 80%",
"rp": "Positivo ++ a +++en el 80%",
"aco": "si", "pid": "091f5a0f-8921-408d-b046-720ae543a072",
"edad": 49, "ki67": "Positivo en mas del 15%",
"peso": "62", "cerb2": "Negativo 1+.",
"hijos": 0, "parto": 0,
"motivo": "Tumor mas dolor", "origen": "FALP",
"cirugia": "Mastectomia total con diseccion axilar reconstruccion expansor.",
"detalle": "moderadamente diferenciado",
"fechacx": "11-27-13", "duracion": "sin hijos",
"familiar": "No tiene antecedentes familiares.",
"lactancia": "sin hijos", "menarquia": 12,
"menopausia": 0, "diagnostico": "Carcinoma ductal"}
```

Secuencias

Atributos

Nombre	secuencias		
Descripción	Registro de resultados de secuenciación de ADN		
Nombre	Tipo de Datos	Descripción	Dominio
id	uuid	Identificador serial	Valor string único positivo
Data	jsonb	data	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
ind	casos	secuencias	1..*	1
consent	consentimientos	secuencias	1	1

Restricciones:

- La data no puede ser nulo.
- El string de individuo debe ser válido y existir en el sistema.
- Debe existir un id de consentimiento válido en el sistema.

Ejemplo data:

```
{"dp": 6648,
"id": "rs7688609",
"vt": "SNP",
"alt": "A", "chr": "chr4", "fep": 0.0029,
"ind": "091f5a0f-8921-408d-b046-720ae543a072",
"nad": "232,367", "ngt": "1/1", "nss": 1, "num": 11, "pos": 1807894, "ref":
"G",
"tad": "164,221", "tgt": "1/1",
"tss": 1,
"gene": "FGFR3", "qual": 12680,
"nfreq": "98.75",
"tfreq": "99.29",
"filter": "PASS",
"secnum": 2, "consent":"2677193f-538b-4a11-94d8-bd70cb147e01"}
```

Consentimientos

Atributos

Nombre	consentimientos		
Descripción	Registro de consentimientos informados		
Nombre	Tipo de Datos	Descripción	Dominio
id	uuid	Identificador serial	Valor string único positivo
Data	jsonb	data	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
ind	casos	consentimientos	1	1

Restricciones:

- La data no puede ser nulo.
- El string de caso debe ser válido y existir en el sistema.
- El tipo puede ser: universal, proyecto, otro por definir.

Ejemplo data:

```
{
"id": “2677193f-538b-4a11-94d8-bd70cb147e01",
 "fecha": “2016-01-23“,
"nombre": “nombre_de_archivo”,
 "tipo": "universal",
“paciente”:”78c055ec-5385-4bc7-a69b-aea4120a9837”,
”proyecto”: {“48894d70-f6be-45a1-9665-08aae18cd77e”,
“457f65d7-3846-4eff-8a52-5de12cdf8930”}
}
```

Subsets

Atributos

Nombre	subsets		
Descripción	Registro de consultas de filtros		
Nombre	Tipo de Datos	Descripción	Dominio
Data	jsonb	data	Array json válido

Asociaciones

Nombre	Clase B	Clase A	Cardinalidad A	Cardinalidad B
proyecto	proyectos	consentimientos	1	1..*

Restricciones:

- La data no puede ser nulo.
- El string de proyecto debe ser válido y existir en el sistema.
- El tipo puede ser: universal, privado.

Ejemplo data:

```
{"id": "2677193f-538b-4a11-94d8-bd70cb147e01",
"fecha": "2016-01-23",
"nombre":
"mujeres mayores de 60",
"tipo": "universal",
"filtro":"AND data->>'sex'='1' AND data->>'edad' >= 60 ",
"proyecto": {"48894d70-f6be-45a1-9665-08aae18cd77e"}
}
```

4.5 Definición de estándares para interface de usuarios

La definición del diseño de las interfaces se estimó de modo que sean cómodas y factibles de utilizar tanto desde dispositivos móviles, tablets en particular, así como desde computadores de escritorio. Este diseño se ajusta automáticamente al tamaño de la pantalla que utilice el usuario.

Se aplicaron los estándares provistos por la plataforma jQuery[1] y Bootstrap[2], tanto en la utilización de elementos de interface como al conjunto de colores y estilos determinados por éste.

4.5.1 Interface

Se definió una barra de menu superior fija que contendrá las opciones de acceso directo para cada una de las funciones definidas del software.

El diseño de la interface para el contenido se basó en el uso de la librería Bootstrap versión 3.2, y utilizando para el control de la interactividad a nivel de usuario principalmente jQuery como herramienta de programación del lado cliente.

4.5.2 Botones y barras de herramientas

Tabla 3. Resumen de elementos de interface gráfica de usuario.

Objeto	Descripción
Etiqueta › Botón estándar para todos los servicios del software.	Edad Campo de formulario para ingreso de datos.
Botón para eliminación de un registro.	Si No Selección de opción mediante botones de tipo check.

[1] http://jquery.com. Librería javascript para control de interactividad del lado cliente.
[2] http://getbootstrap.com. Librería de definición de interface responsiva.

 Botón para selección de acción	 Barra de búsqueda.
 Estándar de botón utilizado para decisiones como Guardar o Cancelar.	 Boton de selección de opciones.

4.6 Diseño detallado del software

4.6.1 Diseño Arquitectónico

La arquitectura del sistema será representada mediante un esquema que evidencia la manera en que el sistema está implementado. Se definen seis áreas de servicios que relacionan el funcionamento de la plataforma propuesta, la colección de datos entrega información al servicio de ETL para cargar la información depurada en la base de datos. El servicio de acceso opera junto con el servicio de seguridad para restringir acceso al sistema. Las entidades con acceso permitido acceden al servicio de referencia para consultas e informes (ver Figura 10).

4.6.2 Diseño de interfaces

El sistema está diseñado para su funcionamiento en dispositivos móviles tipo tablet y equipos de escritorio. Se utilizaron librerías estándares de interfaces basadas en jQuery y Bootstrap.

Se han seguido los lineamientos normalizados para este tipo de dispositivos desplegando los elementos de menú principal en un área superior, y el contenido principal en un área inferior.

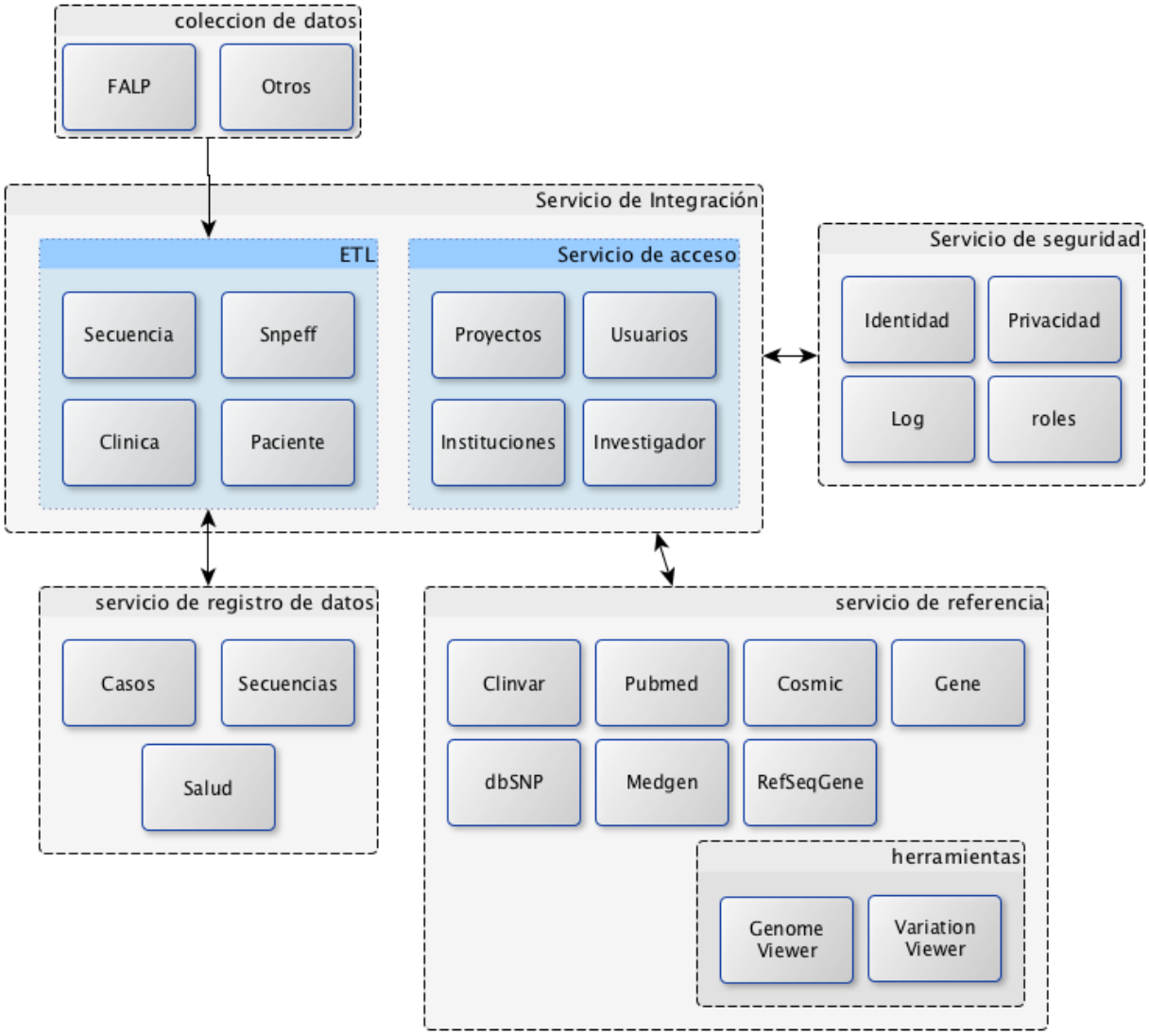

Figura 10. Modelo arquitectónico.

Se indican las áreas de servicio dentro de la plataforma Datagenomed. La colección de datos entrega información a ser integrada y que debe ser depurada por ETL. El servicio de acceso controla los usuarios autentificados y roles dentro del sistema administrado por el servicio de seguridad. El acceso al servicio de referencias también incluye el uso de herramientas bioinformáticas externas a la plataforma. Se registran en la plataforma indormación de casos (diagnósticos), resultados de los procesos de secuenciación de ADN y los antecedentes de salud general que se encuentren disponibles.

4.6.3 Diseño de Pantallas

Acceso principal

La Figura 11 ilustra la forma de acceso al sistema en la direccíón web asignada. El acceso está restringido exclusivamente a usuarios autentificados.

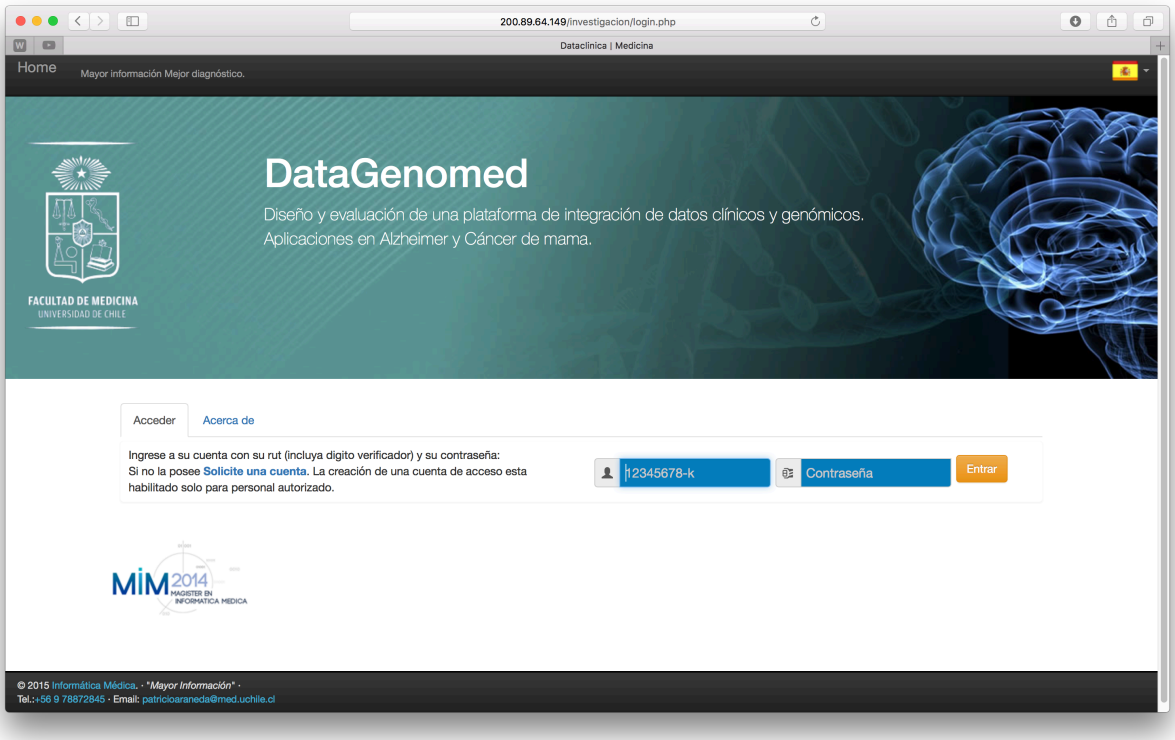

Figura 11. Pantalla de acceso principal. Ingreso de usuario y contraseña.

El acceso a http://assar-lab.cl/investigacion está restringido desde la pantalla de inicio al ingreso de un nombre de usuario (número de rut con dígito verificador) y la contraseña asignada.

Resultados

5 Resultados

La plataforma está actualmente disponible dentro del sitio oficial del Laboratorio Assar-Lab de la Facultad de Medicina, en http://assar-lab.cl/investigacion.

A continuación se explicitan los principales resultados respecto a su construcción, funcionalidad y usabilidad.

Dentro de la definición de la plataforma se generaron una serie de casos de uso para determinar los requisitos necesarios dentro del sistema propuesto.

5.1 Sistema de datos

Respecto del primer objetivo de esta actividad se estableció un modelo de repositorio de naturaleza híbrida para adaptarse a la naturaleza dispar de los datos registrados, permitiendo tanto datos clínicos estructurados como datos genómicos de tipo no estructurados. Se estableció una estructura de datos de tipo jsonb para los registros de la base de datos, debido a la complejidad y variabilidad de la información a contener entre las distintas patologías a analizar. Con este formato de datos fue factible almacenar información tanto de pacientes con Alzheimer como pacientes con diagnóstico de cáncer de mama.

Se decidió utilizar una base de datos PostgreSQL que permite una estructura combinada con un sistema normal relacional para mantener la integridad relacional en aquella información relevante.

Se definió un procedimiento inicial de cargar los datos iniciales, provenientes en su mayoría de planillas Excel, mediante su transformación a formato json, para ser importados a la base de datos definida. Esto se aplicó tanto para los datos médicos como los resultados de la secuenciación de ADN.

Para el caso de Alzheimer se diseñó e implementó una serie de formularios (con su contraparte en la base de datos) para atender las diversas encuestas necesarias para evaluar esta patología. En total sumaron 39 formularios creados a tal fin.

5.1.1 Integración clínico-genómica

La plataforma implementada permite realizar asociación de la información clínica con la información genómica asociada al caso de pacientes específicos mediante la utilización de consultas definidas por el usuario. Los filtros empleados permitieron seleccionar los resultados de secuenciación y la(s) ficha(s) clínica(s) almacenadas en el sistema.

Como forma de consulta se desarrollaron una serie de accesos a bases de datos médicas bibliográficas especializadas como Pubmed y consultas de información genética a bases de datos biológicas dbSNP, Clinvar, Gene, Genmed y Cosmic.

En cuanto a la funcionalidad esperada de la solución Datagenomed, se incorporaron nuevas funciones a nivel de proyectos de investigación para almacenar las consultas realizadas y generar DATASETS de información dándoles un carácter público o privado para ser reutilizadas. Se crearon administradores de grupos, usuarios y proyectos que permitieron flexibilidad en el manejo de información particular a cada proyecto de investigación (también aplicable a instituciones de salud).

Subsets de casos

El resultado del filtrado de casos puede ser almacenado asociado al proyecto específico asignándole un nombre, fecha de modificación, último usuario activo y tipo de disponibilidad, siendo estas últimas las opciones "publico|privado". En caso de ser público estará disponible a todos los usuarios de sistema, en caso de ser privado estará disponible solo para los usuarios que pertenezcan al proyecto indicado en el consentimiento. La disponibilidad estará determinada a su vez por lo siguiente:

- El consentimiento informado sea universal o,
- El consentimiento informado este asociado al proyecto del usuario activo del sistema.

5.1.2 Integración de información

Si bien se estableció en una primera instancia desarrollar una herramienta propia para ETL de los datos involucrados en la plataforma, indicado como parte del objetivo 3, se determinó finalmente de mayor efectividad dado los acotados plazos de tiempo, integrar un software existente a tal efecto. Dentro de las opciones existentes en el mercado se optó por utilizar "Knime"[1] por poseer una amplia red de usuarios y tener conectores a funciones estadísticas en R. Se hicieron pruebas de conexión desde Knime a la base de datos PostgreSQL para extraer y cargar datos.

Se generaron pruebas hacia scripts definidos en R para la generación de reportes estadísticos tal como se muestra en la Figura 12, donde se ejemplifica un reporte de nivel de receptores en el caso de pacientes con cáncer.

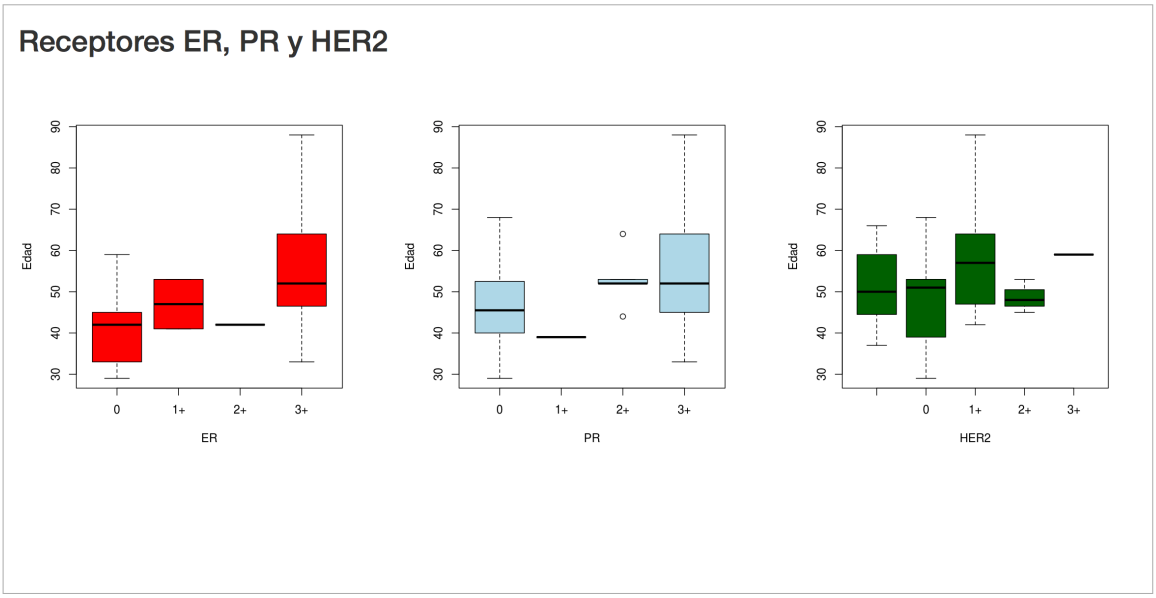

Figura 12. Reporte estadístico generado en R.

Se muestra la distribución de la expresión de receptores en función de la edad. En el gráfico a la izquierda (receptores de estrógeno) se muestra una diferencia significativa al comparar la categoría 3+ respecto a la 0, estando la primera más presente en pacientes de mayor edad (sobre los 50 años).. En el gráfico central (receptores de progesterona) se aprecia que esta diferencia es menos significativa y que las categorías de expresión 1+ y 2+, de varianza reducida, están mayormente diferenciadas.

[1] http://www.knime.org

Respecto de la interface diseñada esta fue aceptada por los usuarios objetivos encontrando buena recepción a su uso. Se utilizó tecnología principalmente javascript para acelerar el rendimiento del lado del cliente en cuanto al manejo de la interface y las conexiones al servidor se restringieron a consultas a la base de datos para renovar información en base a selecciones del usuario.

Lista de Pacientes

La presentación inicial al usuario autentificado es la lista de los pacientes que corresponden al proyecto de investigación al que pertenece. Como se observa en la Figura 13, se visualizan los parámetros de cada registro definidos para el proyecto.

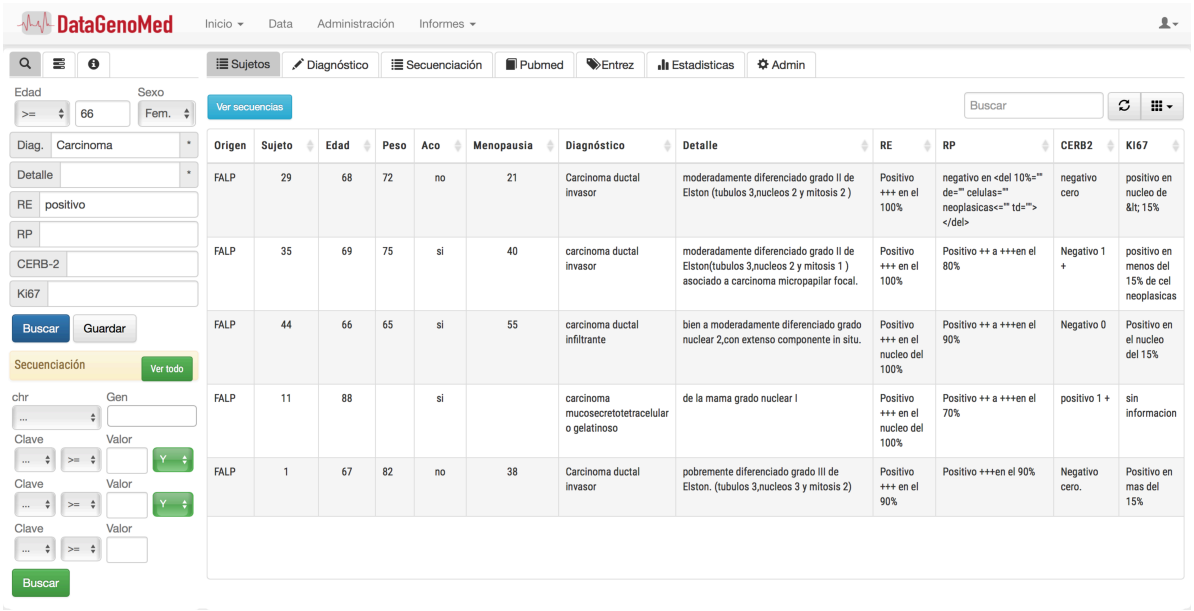

Figura 13. Aplicación de filtros de búsqueda sobre casos de paciente de cáncer.
Se aplicó filtro de edad (mayor o igual a 66 años) y diagnóstico (presencia de carcinoma), mostrando los registros que cumplen las condiciones indicadas.

La información de casos de paciente se despliega en forma de lista, pudiendo ser filtrada para crear subset de casos.

Ficha de diagnóstico/paciente

La ficha de diagnóstico permite ingresar y/o editar la información registrada para cada paciente. Se han incorporado en este formulario la recopilación de la información proporcionada dentro del marco de proyecto “Incorporación de la Secuenciación de Última Generación en el Cuidado de los Pacientes con Cáncer” referente a sujetos de la Fundación Arturo López Peréz (ver Figura 14).

Esta ficha varía en contenido según el diagnóstico asociado al caso.

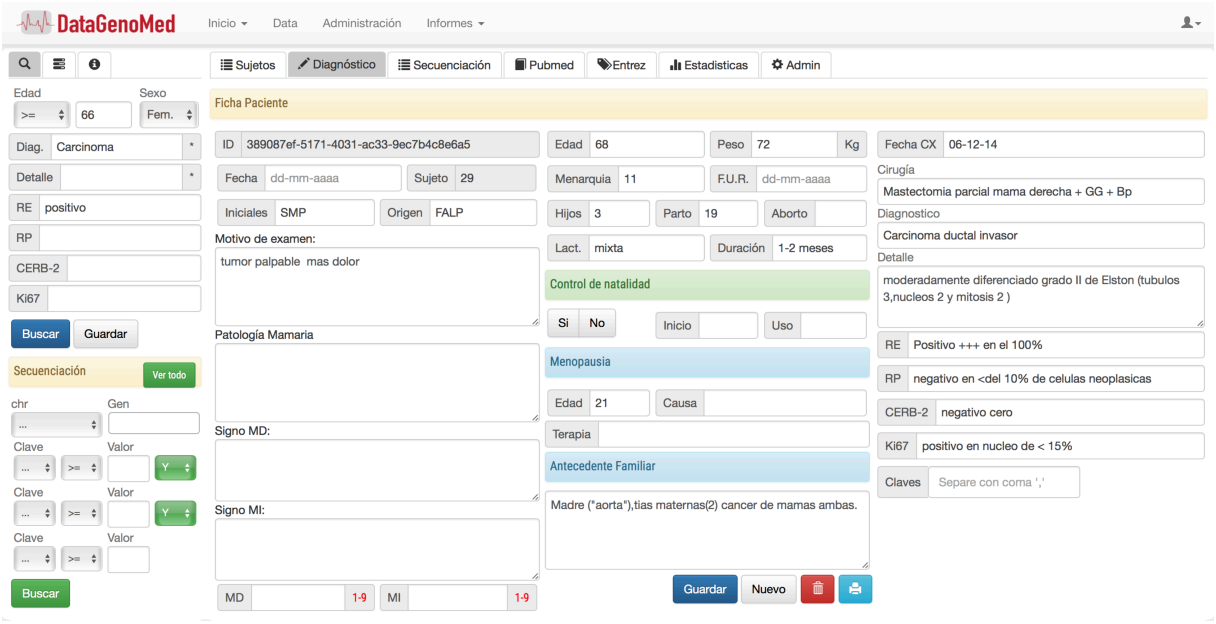

Figura 14. Ficha para visualización de información diagnóstica.

En pacientes con cáncer de mama, la información de la ficha se relacionó con caracterizar el fenotipo de la enfermedad y los antecedentes de salud del paciente. En particular, se incorporó los resultados de receptores de estrógeno, progesterona, CERB-2 y ki67.

En el caso de la información del CA se generaron una serie de formularios de captura de datos de información del paciente (ver Figura 15), donde las interfaces de estos formularios se basó en usabilidad basada mayoritariamente en activación de botones y en menor medida escritura en campos de texto. En total se construyen 39 formularios basados en diversas encuestas en áreas de memoria, comprensión y medición de estrés.

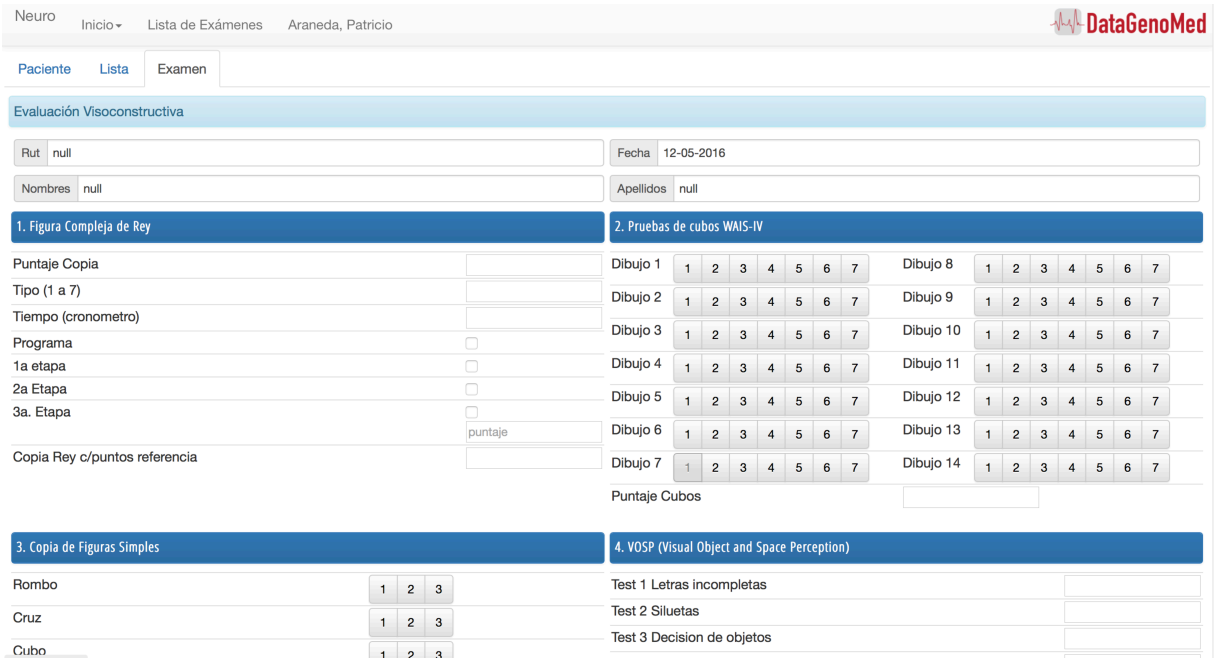

Figura 15. Formulario de encuesta “Evaluación Visoconstructiva”.
Parte de la serie de formularios para la evaluación de pacientes con Alzheimer.

Secuencia

Despliega el resumen de resultados de secuenciación según los casos seleccionados. Estos resultados pueden ser filtrados mediante el filtro de secuenciación. En la Figura 16 se ejemplifica los resultados de secuenciación obtenidos para un paciente específico, donde la columna id y gene se muestran "activas", esto es, se enlazan a las búsquedas hacia las referencias prestablecidas.

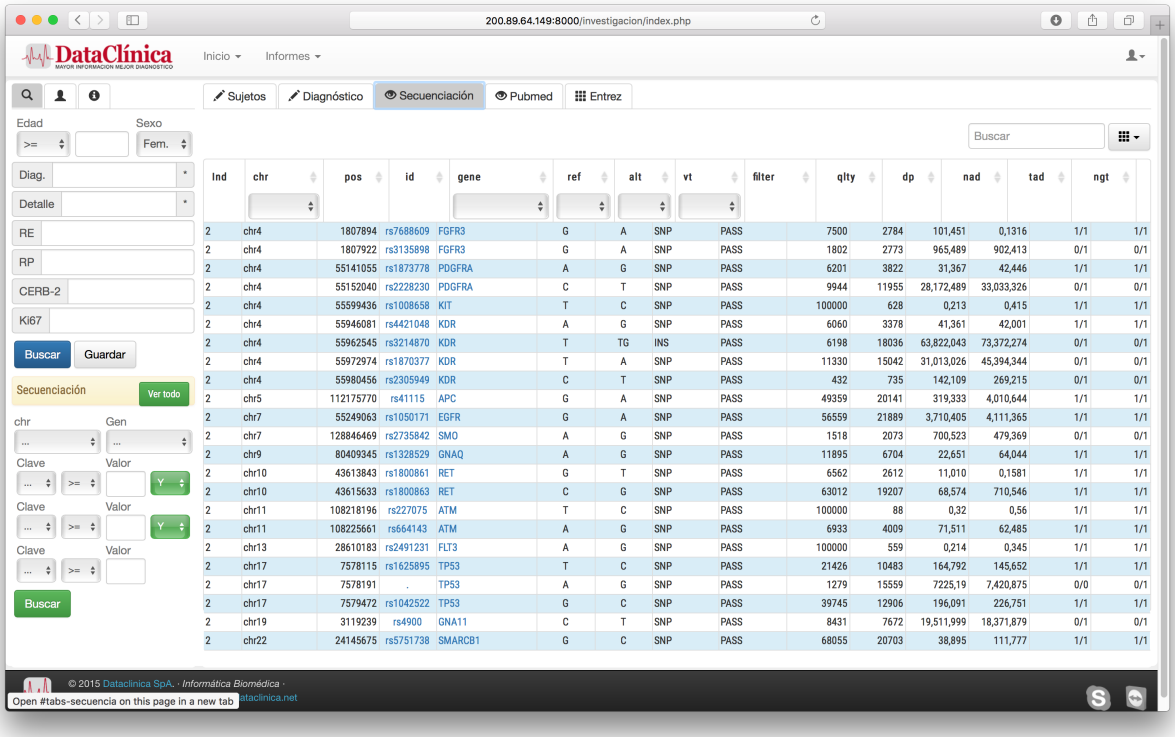

Figura 16. Visualización de variantes genómicas de secuenciación de ADN.
Información resultante del proceso de secuenciación NGS. Las columnas id y gene se muestran activas, indicando que son utilizadas para búsqueda de referencias.

Pubmed

Pubmed[1] es un repositorio de la National Institutes of Health's National Library of Medicine (NIH/NLM), con acceso gratuito a textos completos de artículos de revistas especializadas en biología y medicina. Posee más de cuatro millones de artículos con una participación de cerca de cuatro mil revistas.

El acceso a búsqueda de información bibliográfica en Pubmed sobre datos obtenidos de la secuenciación de ADN despliega una serie de resúmenes de publicaciones y una contabilidad de palabras claves de búsqueda más relevantes como se muestra en la Figura 17. Además permite (si esta disponible) la visualizacion de la publicación en una ventana externa.

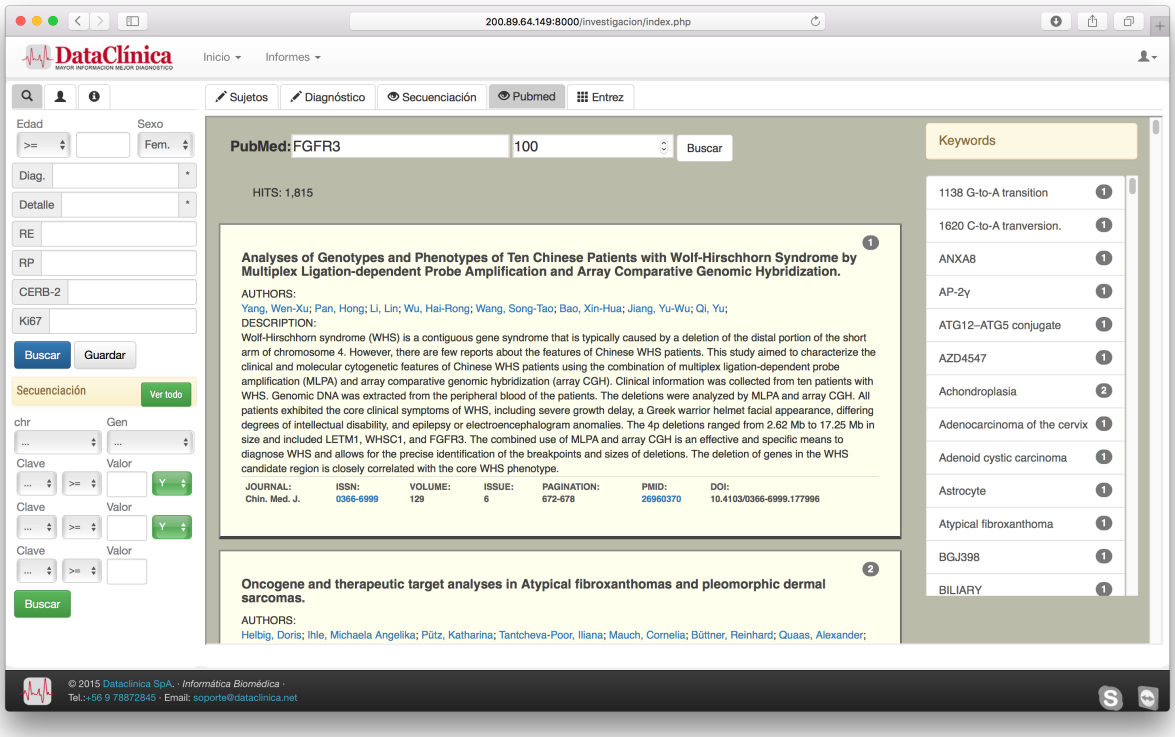

Figura 17. Consulta de información de referencia sobre Pubmed.
Junto con las referencias resultantes que incluyen título del artículo, autor(es) y abstract, se indican las palabras claves y las veces que aparece dentro del término buscado.

[1] https://www.ncbi.nlm.nih.gov/pmc/

Entrez: Gene

Gene[1] Es un repositorio que integra información de varias especies, y otorga conexiones gen-específicas. Incluyen genomas representados por NCBI RefSeqs.

Permite realizar búsquedas de variadas formas: texto libre, nombre de cromosoma y símbolo, nombre de gen, gen con variantes cortas, etc., en la Figura 18 se muestra una búsqueda por nombre de gen (FGFR3).

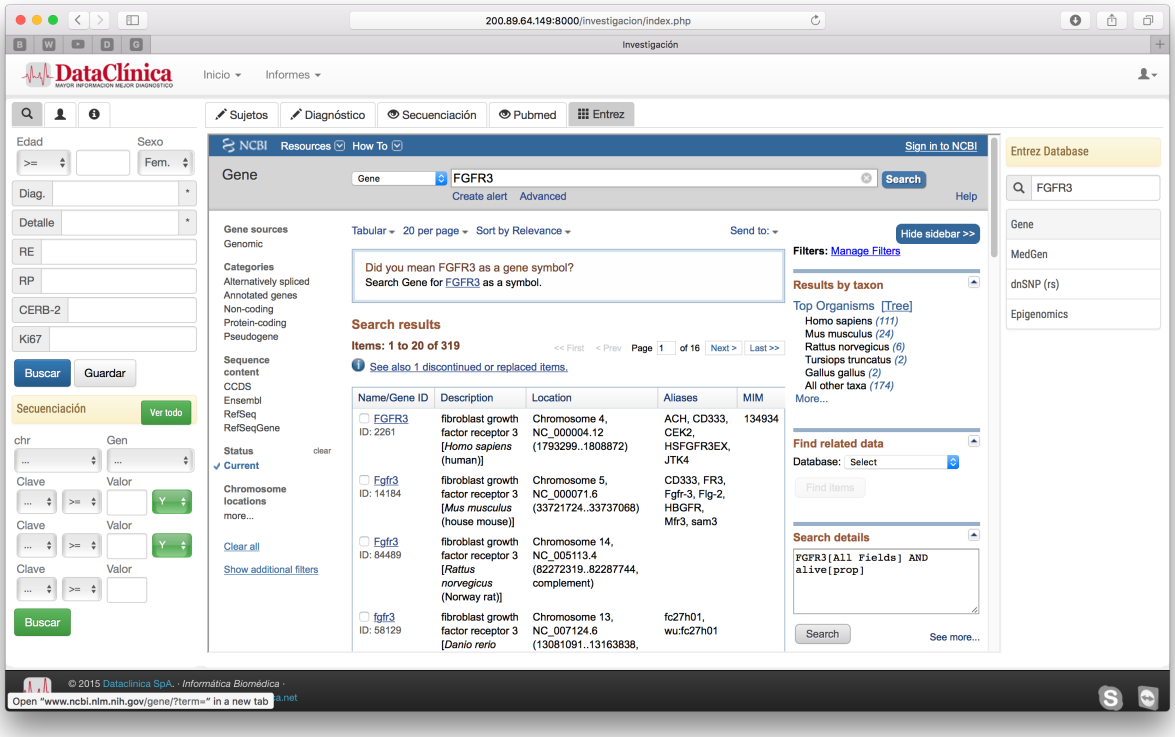

Figura 18. Búsqueda de información en Gene.

Vista de acceso a búsqueda de información bibliográfica en Gene. Búsqueda con el gen FGFR3 en que se obtienen descripciones y locaciones de las repeticiones de dicho gen en distintos cromosomas.

[1] https://www.ncbi.nlm.nih.gov/gene

Entrez: Medgen

Medgen[1] un portal de búsqueda de información relativa a enfermedades y fenotipos que tengan un componente genético. Se integra con otros repositorios como: Clinvar, GTR, MesH

Utiliza un sistema de búsqueda desarrollado por NCBI (Entrez) para buscar términos tanto en Medgen como en otras bases.

En la Figura 19 se muestra el resultado de la consulta por nombre del gen FGFR3 dentro de MedGen.

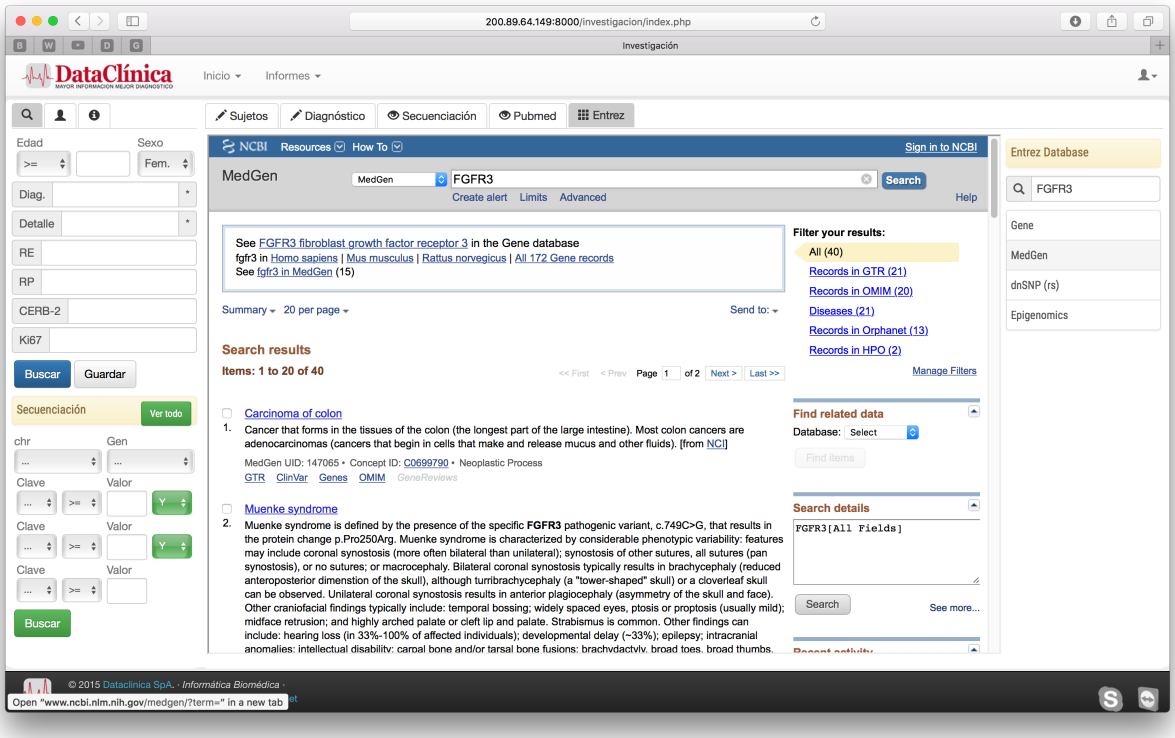

Figura 19. Resultado de búsqueda de gen en Medgen.
Vista de acceso a búsqueda de información bibliográfica en MedGen. Utilizando FGFR3 como patrón de búsqueda se obtiene descripciones de enfermedades asociadas con este gen.

[1] https://www.ncbi.nlm.nih.gov/medgen/

Entrez: dbSNP

La base de datos The National Center for Biotechnology Information (NCBI) Short Genetic Variations conocida popularmente como dbSNP[1], cataloga variaciones en secuencias cortas de nucleótidos SNP (siendo esta la variación genética más común) de diferentes organismos[23]. Las variaciones almacenadas incluyen: nucleótidos simples, inserciones y deleciones, microsatélites y repeticiones en tandem. Posee para algunos casos asociaciones a enfermedades e información genotipica. Creada en 1998, se ha convertido en un repositorio central para la variación genética y cruza información con otras bases como: GenBank, PubMed, LocusLink y otros.

dbSNP posee actualmente 782.342.116 referencias SNP (al 28 de septiembre de 2016).

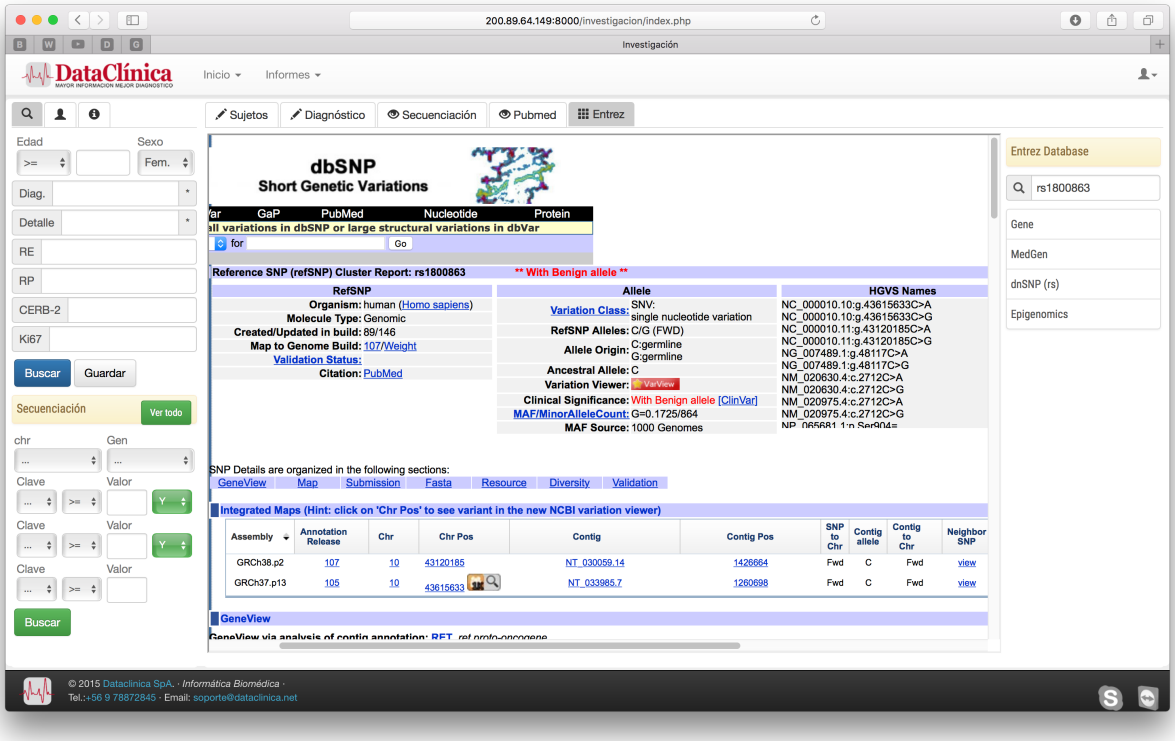

Figura 20. Búsqueda de variante resultante en dSNP.
Provee detalle del registro de variación y coordenadas en cromosoma, junto a consecuencias funcionales de la mutación y la existencia de significancia clínica.

[1] http://www.ncbi.nlm.nih.gov/snp/

Entrez: Cosmic

Cosmic[1] es el más grande catálogo de mutaciones somáticas en cáncer y permite la exploración del impacto de mutaciones en cáncer en humanos. Posee tanto información analizada manualmente por expertos (curada) como información resultante de cargas desde otras fuentes y bases de datos[24].

Las búsquedas en Cosmic se generan por un gen, tipo de cáncer, mutación o por una selección de una región del genoma humano

El acceso a las variables de búsqueda hacia Cosmic se realiza desde la vista de secuenciación o SnpEff. (Figura 21).

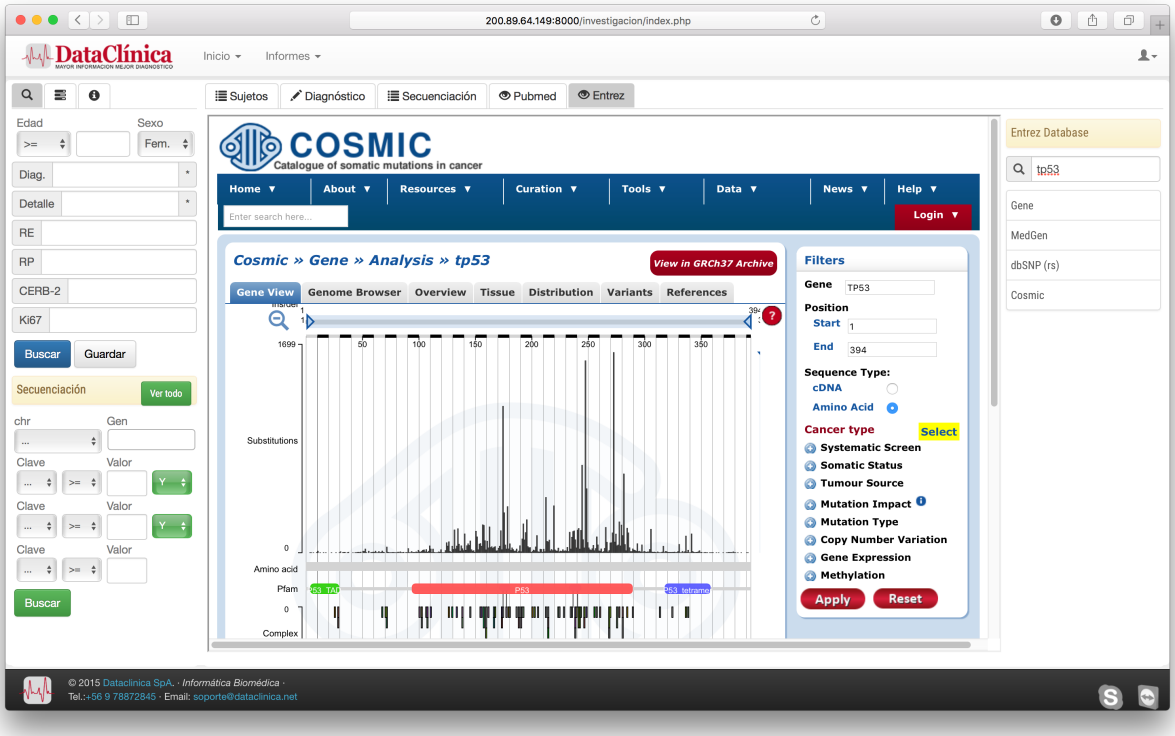

Figura 21. Información de gen en Cosmic.
La selección de gen tp53 arroja la relación de mutaciones indicando el tipo de estas. En esta Figura se alcanza a visualizar la distribución de sustituciones indicando cantidad de mutaciones por posición en el genoma.

[1] http://cancer.sanger.ac.uk/cosmic

Clinvar

Clinvar[1] es un archivo público de acceso gratuito de reportes de relaciones entre variaciones genotípicas en humanos y sus fenotipos basado en evidencia. Incluye el estado de salud y el historial de la observación de las muestras del paciente e información de los remitentes de la información. El nivel de exactitud de las relaciones con la significancia clínica depende de la evidencia que la soporta.

Este repositorio permite búsquedas de términos como símbolos de genes: PTEN, Locación en un cromosoma: 10[chr] AND 89623000:89730000[chrpos37] o Números rs: rs180177042 (ver Figura 22).

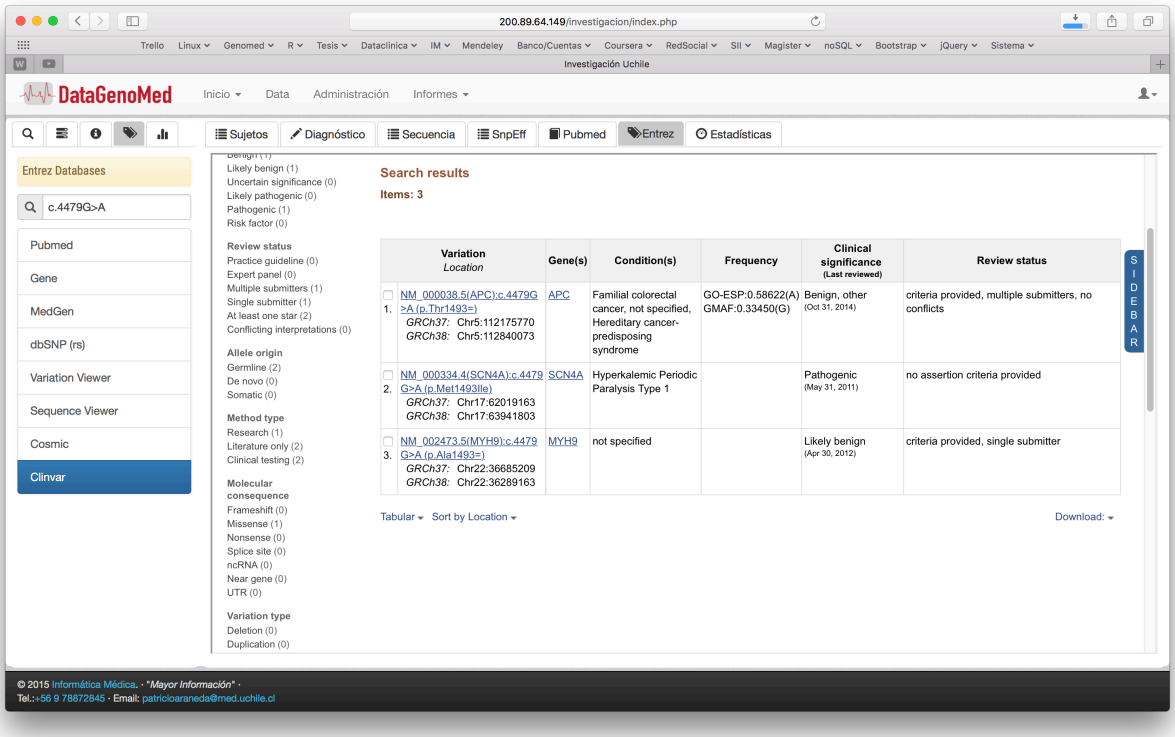

Figura 22. Resultado de búsqueda en Clinvar.
La búsqueda de locación de variante arroja locación, asociaciones con enfermedades, el nivel de significancia clínica (patógeno o benigno).

[1] http://www.ncbi.nlm.nih.gov/clinvar/

RefSeqGene

RefSeqGene[1] define secuencias a ser usadas como estándar sobre genes bien caracterizados y sirven para reportar mutaciones y definir secuencias coordinadas de variaciones genéticas. Enlaza además con Gene, Medgen y Clinvar y otras bases.

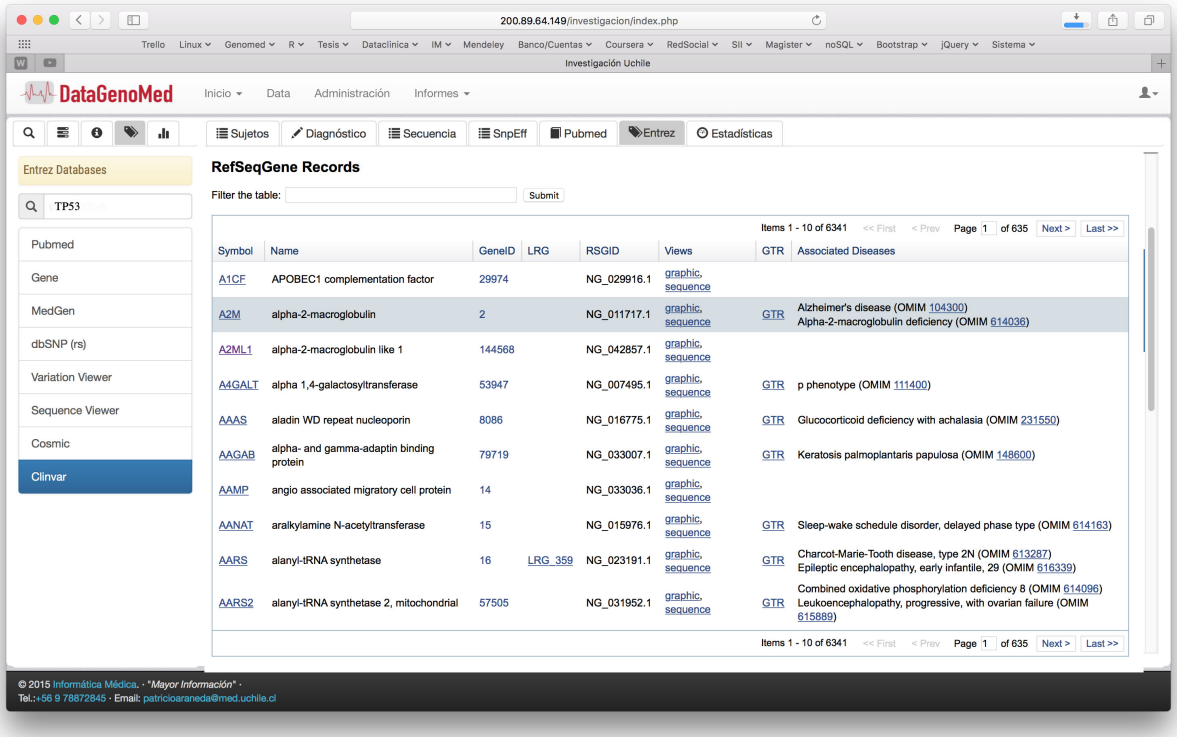

Figura 23. RefSeqGene

La búsqueda de gen TP53 arroja descripción de enfermedades asociadas y conecta a la base Nucleotide para visualizar la secuencia completa.

[1] https://www.ncbi.nlm.nih.gov/refseq/rsg/

Administración

El área de administración permite gestionar el control de usuarios, roles y pertenencia a grupos. Cada usuario puede pertenecer a más de un grupo de investigación o institución de salud. El acceso a la información también esta definido por el rol del usuario (ver Figura 24).

Los resultados que son indicados como privados, solo pueden ser visualizados por los usuarios asociados al proyecto de investigación que genera/administra esos datos.

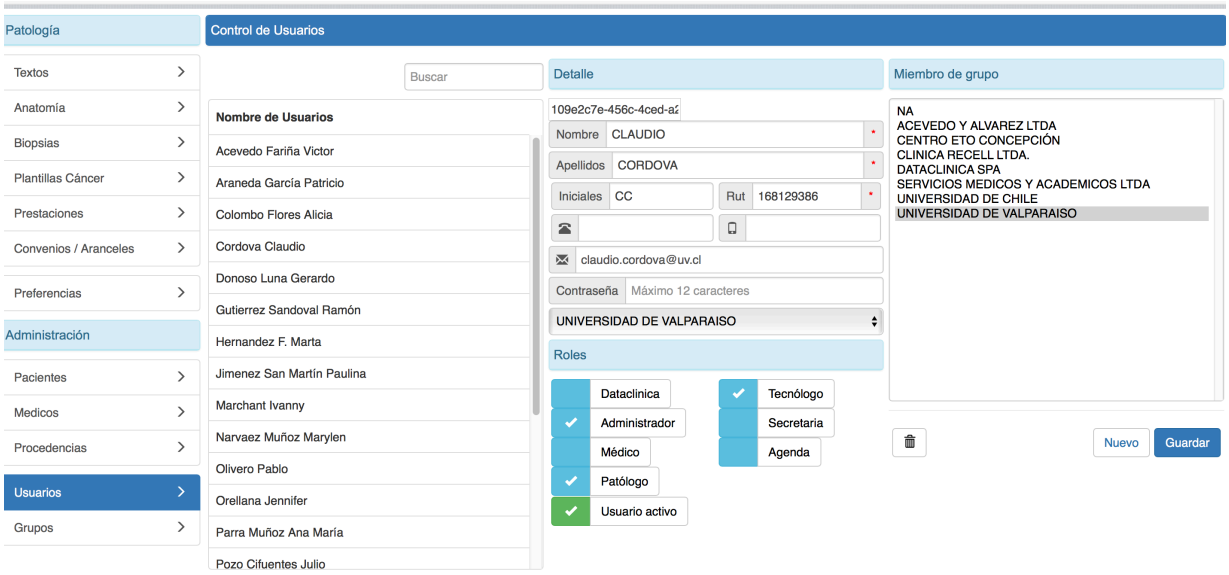

Figura 24. Administración de usuarios y grupos.

La administración de usuarios y grupos es realizada únicamente por el administrador de sistema.

5.2 Usabilidad

La evaluación de la encuesta de usabilidad que se aplicó a los usuarios del sistema (se encuestaron a los cuatro usuarios), arrojó una buena recepción de parte de ellos hacia la funcionalidad general del sotware.

Las mejores evaluaciones se obtuvieron en el ámbito de uso de terminología (comunicación: plataforma-->usuario) con un promedio de 7.8 puntos sobre un máximo de 9, y en capacidades del sistema con 7.4 sobre 9.

En la Figura 25 se visualiza la evaluación usuaria respecto del uso de terminologías medidas en una escala de 0 de 9 (ver encuesta en anexo 1). Con una alta evaluación respecto de los mensajes en cuanto a ubicación y retención. La menor valoración se encuentra en exploración y aprendizaje lo que podría indicar la necesidad de una capacitación previa al uso de la plataforma.

En Figura 26 y Figura 27 se expresan los resultados negativos y positivos por los encuestados respecto a la manipulación de la plataforma (comunicación Usuario-->Plataforma). En Figura 26 (donde menor valor es mejor), sólo uno de los encuestados encontró difícil utilizar la plataforma, en tanto el resto no encontró dificultades en su uso

El resultado mostrado en la Figura 27 indica alta valoración general del uso de la plataforma, donde obtuvieron mejores puntuaciones los temas de facilidad, control y atractivo.

En el aspecto de usabilidad, la plataforma creada resultó ser sencilla de utilizar por los usuarios, la comunicación desde plataforma-->usuario fue bien evaluada.

Uso de terminología

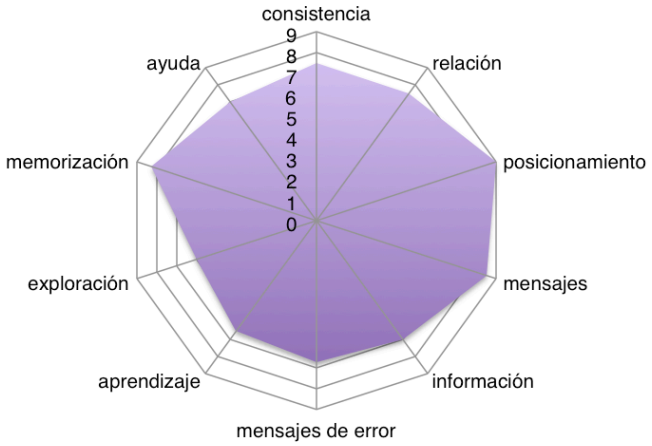

Figura 25. Evaluación usuaria de plataforma. Uso de terminología.
Los mejores evaluaciones (Mayor valor es mejor) fueron en relación a la posición y generación de los mensajes emitidos al usuario con 9 y 8,5 puntos respectivamente, y la capacidad de retención con 8,25 puntos. El resultado en exploración y aprendizaje sugiere la necesidad de capacitación para el mejor uso de la plataforma.

Evaluación usuaria – aspectos negativos

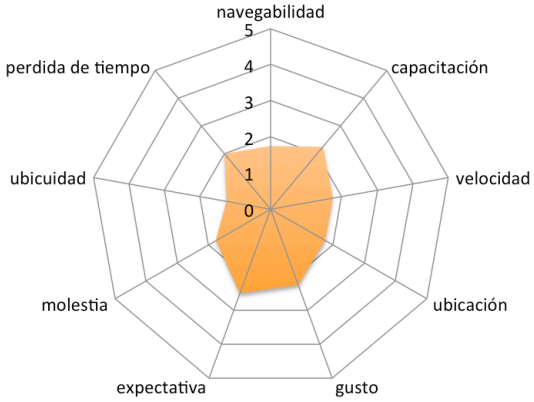

Figura 26. Evaluación usuaria de la plataforma Datagenomed.
Visualización de aspectos considerados negativos (donde menor valor es mejor).

Evaluación usuaria – aspectos negativos

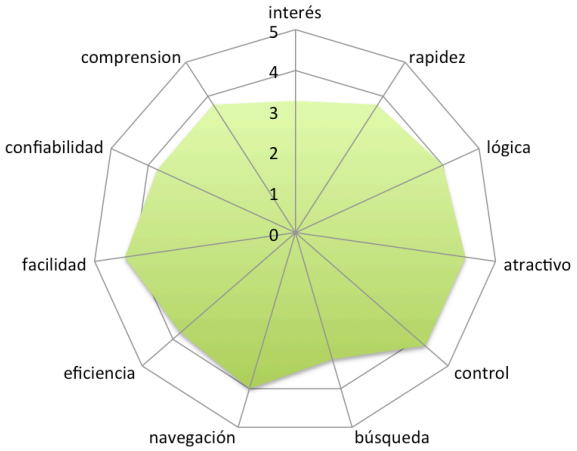

Figura 27. Evaluación usuaria respecto de aspectos positivos.
Los temas mejor valorados (mayor valor es mejor) fueron facilidad de uso, control y atractivo con 4,25 puntos de un total de 5. Las áreas más débiles fueron interés y búsqueda con 3,25.

Conclusión y Discusión

6 Conclusiones

Se implementó un prototipo de plataforma que permite asociar la información genómica y clínica de un paciente y presentar esa información de una manera simple para visualizar y realizar consultas hacia fuentes de información externas. Esta plataforma simplifica en gran medida el trabajo de investigación genotípica al unificar las fuentes de información bajo una misma herramienta. Al respecto la evaluación del sistema en base a la encuesta aplicada a los usuarios arroja una alta valoración respecto de su utilidad y facilidad de uso.

El acceso oportuno y eficiente a las fuentes de información a través de esta plataforma permite descubrir relaciones documentadas de variantes genómicas o mutaciones, con enfermedades y/o tratamientos eficientes o asociaciones con fármacos, sobre todo como se muestra en las figuras 21 y 22 cuando se utilizan fuentes específicas como Cosmic y Clinvar.

La elección de la estructura de datos seleccionada, específicamente postgreSQL (con un modelo NoSQL), resultó ser efectiva en el almacenamiento de información de los distintos dominios, información biomédica que se caracteriza por ser heterogénea según la patología a estudiar. Permitió una mayor simplicidad en la adecuación de los registros y las modificaciones permanentes que se realizan en una base de datos de este tipo y facilitó el desarrollo del software al unificar las funciones de manejo de los datos. Sin embargo, con esta estructura se debió atender con mayor prolijidad el tema de consistencia de datos dentro del software.

Respecto de los objetivos planteados para esta AFE, se completó la implementación de un repositorio de datos con capacidad de aglutinar la información clínica y genómica, parte del primer objetivo, y el desarrollo de la plataforma software que manipula estos registros para ser consultados y contrastados con información externa, definido en el segundo objetivo.

En base a lo anterior y a la correcta asociación de la información clínica y genómica al seleccionar un caso/paciente determinamos que se cumplió con el objetivo general de creación de una plataforma de integración de estas áreas, en los casos analizados.

La construcción de interfaces de ETL se alteró hacia la utilización de una software ya construido y probado a este fin (Knime) por la complejidad y alcance de tiempo para desarrollar un software

para esta finalidad. También se probó la viabilidad de utilizar R directamente desde la plataforma Datagenomed para la generación de reportes estadísticos simples, lo que se corroboró en la emisión de reporte mostrado en la Figura12, informe de nivel de receptores en función de la edad en pacientes con cáncer de mama.

7 Discusión

Esta plataforma permitió realizar las consultas de referencia hacia bases bibliográficas y bases de datos biomédicas tanto en NCBI y otras fuentes. Las referencias pueden aumentarse a solicitud del investigador según el objetivo de la investigación. Estas consultas se establecieron mediante accesos directos a la url de búsqueda según documentación de cada repositorio. Una segunda etapa debería consistir en utilizar o desarrollar APIs[1] para acceder a la información de las fuentes externas, como también crear una serie de webservices o funciones REST[2] para la consulta efectiva de resultados hacia y desde otros sistemas informáticos.

Toda la plataforma se centralizó en un solo servidor, tanto para la base de datos como la aplicación, además de la información secuencial de ADN y el servicio R. Sin embargo, es factible y más eficiente implementar esta plataforma en un cluster, de forma que las llamadas a rutinas R se realicen hacia un servidor potenciado a tal efecto y en otro servidor estén almacenadas los archivos Fastq para ser procesados.

La interconexión probada con R permitiría desde esta plataforma, generar y/o procesar pipelines de análisis de información directamente desde la interface web, previa selección del conjunto de datos y de los procesos a realizar. Además, el archivo Fastq resultante de la secuenciación puede ser accedido desde la plataforma, asociarlo al paciente y ser reprocesado mediante la selección de rutinas de R.

[1] Application Programming Interface (API), Interface de comunicación hacia programas externos.

[2] Transferencia de Estado Representacional (REST) estilo de arquitectura para sistemas distribuidos, define operaciones entre sistemas generalmente sobre Internet.

Dentro de las tareas a implementar para automatizar el proceso extracción y carga de datos se encuentra la ampliación de estrategias de ETL mediante el eso del programa "Knime" para la realización de tareas de extracción y limpieza de datos sobre los archivos VCF y bases de datos del sistema y mayores pruebas sobre datos clínicos externos de distinta procedencia.

Resumiendo, en posteriores etapas (asociadas a proyectos postulados) se espera:

- Elaborar y seleccionar métodos y herramientas para permitir el análisis estadístico integrativo y predictivo mediante interfaces con programa R.
- Evaluación de la calidad de los datos en el data warehouse .
- Establecer conexiones mediante API hacia repositorios que cuentan con soporte REST.
- Evaluación de resultados de prototipo Datagenomed: calidad de resultados, análisis e interface.
- Establecer procesos de mejora.

Siendo esta una solución destinada a cubrir una necesidad específica de análisis genómico basado en secuenciación de ADN y variantes genéticas, queda probar si es factible de incorporar información de tipo ómica (por ejemplo Transcriptómica) e incorporar además a otras entidades externas como laboratorios de patología y/o biobancos respecto de la captura de información adicional, tanto diagnóstica como de muestras clínicas y de esa manera conformar una red de investigación. En Figura 8 se esquematiza una propuesta de integración entre instituciones que puede generar en una plataforma unificada para procesamiento y reportabilidad de investigación genómica.

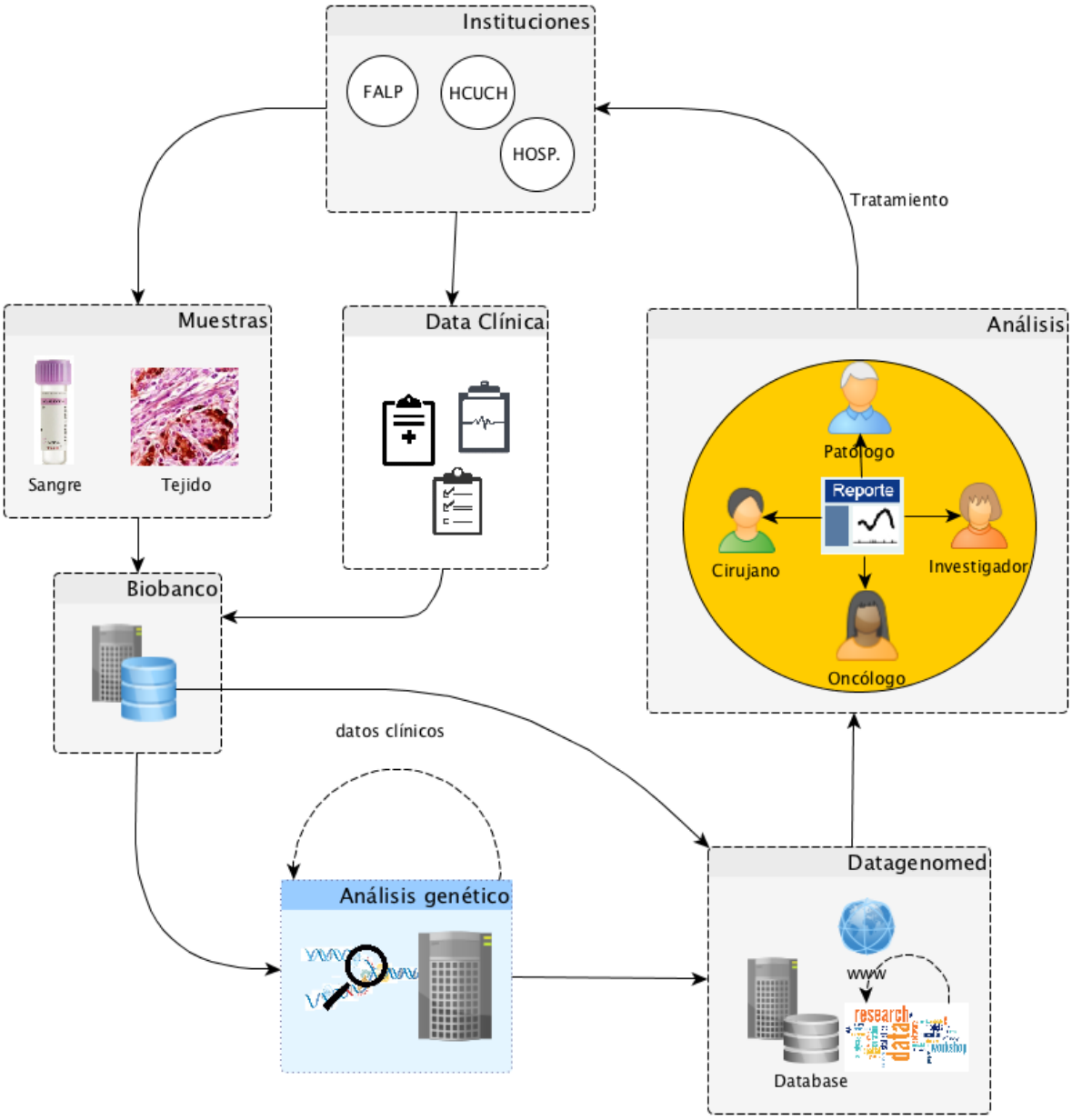

Figura 28. Propuesta de integración Datagenomed.

Las muestras obtenidas para investigación junto a la data clínica se concentra en un repositorio central definido en un biobanco de tejidos y datos. Desde este punto se recojen las muestras para análisis genético (u otro tipo) y los resultados se cargan en Datagenomed para ser consultados por investigadores y/o personal médico.

Bibliografía

8 Bibliografía

[1] R. Sabatier, A. Gonçalves, and F. Bertucci, "Personalized medicine: present and future of breast cancer management.," *Crit. Rev. Oncol. Hematol.*, vol. 91, no. 3, pp. 223–33, Sep. 2014.

[2] L. J. Lesko, "Personalized medicine: elusive dream or imminent reality?," *Clin. Pharmacol. Ther.*, vol. 81, no. 6, pp. 807–816, 2007.

[3] M. G. de Lecea and M. Rossbach, "Translational genomics in personalized medicine – scientific challenges en route to clinical practice," *Hugo J.*, vol. 6, no. 1, p. 2, 2012.

[4] O. E.Sheta, "Building A Health Care Data Warehouse for Cancer Diseases," *Int. J. Database Manag. Syst.*, vol. 4, no. 5, pp. 39–46, Oct. 2012.

[5] J. C. Denny, "Surveying Recent Themes in Translational Bioinformatics: Big Data in EHRs, Omics for Drugs, and Personal Genomics.," *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 199–205, Jan. 2014.

[6] L. J. Frey, L. Lenert, and G. Lopez-Campos, "EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group.," *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 206–11, Jan. 2014.

[7] R. L. Richesson, M. M. Horvath, and S. a Rusincovitch, "Clinical research informatics and electronic health record data.," *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 215–23, Jan. 2014.

[8] V. A. Fusaro, P. Patil, E. Gafni, D. P. Wall, and P. J. Tonellato, "Biomedical cloud computing with Amazon Web Services.," *PLoS Comput. Biol.*, vol. 7, no. 8, p. e1002147, Aug. 2011.

[9] A. Abbas and S. U. Khan, "A review on the state-of-the-art privacy-preserving approaches in the e-health clouds.," *IEEE J. Biomed. Heal. informatics*, vol. 18, no. 4, pp. 1431–41, Jul. 2014.

[10] B. Knoppers, "Framework for responsible sharing of genomic and health-related data," *Hugo J.*, vol. 8, no. 1, p. 3, 2014.

[11] V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun, "Translational research platforms integrating clinical and omics data: a review of publicly available solutions.," *Brief. Bioinform.*, Mar. 2014.

[12] D. Segagni, V. Tibollo, A. Dagliati, A. Zambelli, S. G. Priori, and R. Bellazzi, "An ICT infrastructure to integrate clinical and molecular data in oncology research.," *BMC Bioinformatics*, vol. 13 Suppl 4, no. Suppl 4, p. S5, Jan. 2012.

[13] M. D. Natter, J. Quan, D. M. Ortiz, A. Bousvaros, N. T. Ilowite, C. J. Inman, K. Marsolo, A. J. McMurry, C. I. Sandborg, L. E. Schanberg, C. a Wallace, R. W. Warren, G. M. Weber, and K. D. Mandl, "An i2b2-based, generalizable, open source, self-scaling chronic disease registry.," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 172–9, Jan. 2013.

[14] J. G. Klann and S. N. Murphy, "Computing health quality measures using Informatics for Integrating Biology and the Bedside.," *J. Med. Internet Res.*, vol. 15, no. 4, 2013.

[15] V. G. Deshmukh, S. M. Meystre, and J. A. Mitchell, "Evaluating the informatics for integrating biology and the bedside system for clinical research.," *BMC Med. Res. Methodol.*, vol. 9, p. 70, 2009.

[16] D. Segagni, V. Tibollo, A. Dagliati, A. Malovini, A. Zambelli, C. Napolitano, S. G. Priori, and R.

Bellazzi, “Clinical and research data integration: the i2b2-FSM experience.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2013, pp. 239–40, 2013.

[17] D. Segagni, V. Tibollo, A. Dagliati, L. Perinati, A. Zambelli, S. Priori, and R. Bellazzi, “The ONCO-I2b2 project: Integrating biobank information and clinical data to support translational research in oncology,” in *Studies in Health Technology and Informatics*, 2011, vol. 169, pp. 887–891.

[18] A. J. McMurry, S. N. Murphy, D. MacFadden, G. Weber, W. W. Simons, J. Orechia, J. Bickel, N. Wattanasin, C. Gilbert, P. Trevvett, S. Churchill, and I. S. Kohane, “SHRINE: enabling nationally scalable multi-site disease studies.,” *PLoS One*, vol. 8, no. 3, p. e55811, Jan. 2013.

[19] C. Safran, “Reuse of clinical data.,” *Yearb. Med. Inform.*, vol. 9, no. 1, pp. 52–4, Jan. 2014.

[20] V. Abramova and J. Bernardino, “NoSQL databases,” in *Proceedings of the International C* Conference on Computer Science and Software Engineering - C3S2E ’13*, 2013, pp. 14–22.

[21] G. Lindgaard and J. Chattratichart, “Usability testing,” in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’07*, 2007, p. 1415.

[22] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden, “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff,” *Fly (Austin).*, vol. 6, no. 2, pp. 80–92, Apr. 2012.

[23] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, “dbSNP: the NCBI database of genetic variation.,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–11, Jan. 2001.

[24] S. A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, and M. R. Stratton, “The Catalogue of Somatic Mutations in Cancer (COSMIC).,” *Curr. Protoc. Hum. Genet.*, vol. Chapter 10, p. Unit 10.11, Apr. 2008.

[25] L. Barillot, Emmanuel; Hupé, Philippe; Calzone, *Computational Systems Biology of Cancer*. CRC Press, 2013.

Anexo 1

Encuesta de Usabilidad

Encuesta de usabilidad

Por favor, compruebe que ha dado una respuesta a cada punto. Caso de no estar seguro acerca de la evaluación de alguno de los mismos, elija, por favor, la columna intermedia.
La información que proporciona se mantiene de forma completamente confidencial.

Reacciones globales

Terrible	0	1	2	3	4	5	6	7	8	9	Bueno
Difícil	0	1	2	3	4	5	6	7	8	9	Fácil
Frustrante	0	1	2	3	4	5	6	7	8	9	Satisfactorio
Potencia adecuada	0	1	2	3	4	5	6	7	8	9	Potencia inadecuada
Aburrido	0	1	2	3	4	5	6	7	8	9	Estimulante
Rígido	0	1	2	3	4	5	6	7	8	9	Flexible

Terminología

1. Uso de terminología en el sistema										
Inconsistente	0	1	2	3	4	5	6	7	8	9	Consistente
2. La terminología se relaciona con la tarea										
Nunca	0	1	2	3	4	5	6	7	8	9	Siempre
3. Posición de los mensajes en pantalla										
Inconsistente	0	1	2	3	4	5	6	7	8	9	Consistente
4. Mensajes en pantalla que indican al usuario que introduzca datos										
Confuso	0	1	2	3	4	5	6	7	8	9	Muy claro
5. El sistema mantiene informado al usuario de lo que está sucediendo										
Nunca	0	1	2	3	4	5	6	7	8	9	Siempre
6. Mensajes de error										
No ayudan	0	1	2	3	4	5	6	7	8	9	De gran ayuda
7. Aprendiendo a operar el sistema										
Difícil	0	1	2	3	4	5	6	7	8	9	Fácil
8. Exploración de nuevos aspectos por ensayo y error										
Difícil	0	1	2	3	4	5	6	7	8	9	Fácil
9. Memorización de nombres y uso de comandos										
Difícil	0	1	2	3	4	5	6	7	8	9	Fácil
10. Mensajes de ayuda en pantalla										
No ayudan	0	1	2	3	4	5	6	7	8	9	De gran ayuda

Capacidades del sistema

1. Velocidad del sistema

Demasiado lento	0	1	2	3	4	5	6	7	8	9	Muy rápido
2. Fiabilidad del sistema											
No es fiable	0	1	2	3	4	5	6	7	8	9	My fiable
3. Corrección de errores propios											
Difícil	0	1	2	3	4	5	6	7	8	9	Fácil
4. Medida en que se tiene en cuenta a los usuarios sin experiencia											
Nunca	0	1	2	3	4	5	6	7	8	9	Siempre

CUESTIONARIO

La parte principal de este cuestionario consiste en 20 puntos. Considérelos, por favor, de forma cuidadosa y evalúe su acuerdo con cada uno de los mismos haciendo uso de la escala de cinco puntos desde completamente de acuerdo hasta en completo desacuerdo.

¿a ples fueron sus principales razones para hacer uso de este sitio web?:

.	Completo de acuerdo Completo desacuerdo				
Este sitio web tiene muchas cosas de interés para mí					
Es difícil moverse por este sitio web					
Puede encontrar rápidamente lo que quiero en este sitio web					
Este sitio web me parece bastante lógico					
Este sitio web necesita más explicaciones introductorias					
Las páginas de este sitio web son muy atractivas					
Tengo el control cuando me muevo por este sitio web					
Este sitio web es demasiado lento					
Este sitio web me ayuda a encontrar lo que busco					
Situarme en este sitio web es un problema					
La navegación entre las páginas es fácil					
No me gusta utilizar este sitio web					
Me siento eficiente al utilizar este sitio web					

Es difícil decir si este sitio web tiene lo que quiero					
Es fácil utilizar este sitio web por primera vez					
Este sitio web tiene algunas características molestas					
Es difícil tratar de recordar donde estoy en este sitio web					
Es una pérdida de tiempo usar este sitio web					
Sale lo que espero cuando sigo los vínculos en este sitio web					
En este sitio web, todo es fácil de entender					

FUNCIONALIDAD

Considere el siguiente caso de uso del sistema y conteste las preguntas:

Cáncer de mama: Objetivo. Seleccione un conjunto de pacientes en base a una serie de criterios por Ud. Elegido (puede elegir mas de uno) y revisar la información bibliográfica de una mutación de interés.

Tareas (NR: no realizada, MT: medio terminar, TC: terminada correctamente	NR	MT	TC
Ingreso a pantalla de lista de pacientes			
Selección de criterios y Filtrado de pacientes			
Consultar diagnostico de un paciente			
Listar los resultados de secuenciación de todos los pacientes elegidos			
Filtrar información de secuenciación por criterios elegidos por Ud.			
Visualizar información bibliográfica de un gen en particular			
Visualizar información referencial de un snp			

Alzheimer: Objetivo. Definir resultados de depresion y/o memoria según los resultados de una serie de encuetas.

Tareas (NR: no realizada, MT: medio terminar, TC: terminada correctamente	NR	MT	TC
Seleccionar un paciente del total registrado			
Seleccionar los resultados de las encuestas definidas para el paciente			

Determinar score especifico			

Existe un caso de uso que desee agregar? Cuál (especifique las tareas requeridas)
