Informes genéticos

José Luis Martínez Pérez
Máster universitario de Bioinformática y Bioestadística
Área 12: Estudios genéticos de enfermedades humanas

Helena Brunel
David Merino Arranz

19 de Marzo de 2018.

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Informes genéticos*
Nombre del autor:	*José Luis Martínez Pérez*
Nombre del consultor/a:	*Helena Brunel*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	06/2018
Titulación::	*Máster Bioinformática y Bioestadística*
Área del Trabajo Final:	*Estudios genéticos de enfermedades humanas*
Idioma del trabajo:	Español
Palabras clave	*Bioinformática, medicina personalizada, SNPs.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El propósito de este trabajo es la realización de una plataforma web que sea capaz de analizar datos SNPs y ofrecer informes genéticos relacionados con los datos SNPs.

Los objetivos principales son el estudio de cómo realizar el diseño e implementación de este tipo de plataformas y por otra parte, profundizar en las bases de datos ómicas existentes actualmente.

Como conclusiones más importantes es conveniente destacar la heterogeneidad de los datos ómicos, así como las diferentes bases de datos existentes y en muchas ocasiones, la dificultad para encontrar e interpretar la información que se necesita.

Abstract (in English, 250 words or less):

Bioreports is an application capable of analyze SNPs data and offer as a result several genetics reports related to SNPs data.

The main objectives are the study and evaluation of how to design and implement this sort of applications and also, look in depth the current omic databases.

As a conclusion, it is important to say the diversity of the omic data, depending on which database or source you get them, the data can be presented in several ways. That diversity can cause difficulties in order to interpret and understand the data.

Índice

Lista de figuras

No se encuentran elementos de tabla de ilustraciones.

1. Introducción

La **Medicina Genómica** es la ciencia que aplica los conocimientos derivados de la descodificación del genoma humano para la **predicción de los riesgos individuales a padecer determinadas patologías**.

Estudia diversos cambios en el genoma de cada persona, que directamente y aisladamente no le producen ninguna enfermedad, pero que en conjunto pueden predisponerle a padecerlas.

La medicina personalizada ó genómica es sin duda una de las disciplinas científicas con más auge actualmente y en la que más énfasis se está poniendo por parte de los investigadores actualmente.

Tiene como objetivo identificar marcadores biológicos para detectar a los individuos con alto riesgo de padecer determinadas afecciones. El conocimiento a priori de las enfermedades que puede padecer una persona dada su genética presenta unos beneficios a distintos niveles:

- Social. Teniendo consciencia de la predisposición genética a contraer y/o padecer ciertas enfermedades y qué pautas de prevención o tratamiento se ajustan mejor a cada organismo mejora indudablemente la calidad de vida de las personas y su esperanza de vida.
  Económico. Los tratamientos de prevención de enfermedades son más baratos que los tratamientos de la enfermedad una vez contraída. Por ejemplo, cánceres que pueden prevenirse o detectarse precozmente, hipertensión, enfermedades cardiovasculares, diabetes, etc. Por lo tanto, las medicina personalizada permite ahorrar en tratamientos y medicamentos al sistema de salud.

1.1 Contexto y justificación del trabajo

Dentro del contexto presentado en el punto anterior, este trabajo se justifica debido a la necesidad de continuar trabajando en la mejora de los procedimientos y software que se utilizan en la medicina genómica así como aportar conocimiento sobre las tecnologías y ciencias aplicadas en ella.

1.1.1 Descripción general

Actualmente existen empresas que realizan informes genéticos, tanto a nivel nacional como internacional. Estas empresas, suelen recoger una muestra de ADN del usuario, la analizan y le entregan sus distintos informes genéticos de

forma digital, a través de su plataforma online e impresa en papel. El proceso de análisis es complejo y proporciona información muy variada dependiendo de tipo de prueba que se contrate. Informes sobre enfermedades, nutrición y ancestros son algunos ejemplos de informes que estas empresas pueden proporcionar a sus clientes.
El proyecto va a emular una plataforma similar a las ya existentes con el objetivo de dar a conocer tales plataformas y la metodología y procesos bioinformáticos que hay detrás de estos informes genéticos.

1.1.2 Justificación del TFM

A través de este TFM, se pretende dar a conocer los procedimientos internos del proceso de generación de informes genéticos. El resultado principal del TFM será el desarrollo de una plataforma similar a las existentes junto con una explicación detallada de los subprocesos que intervienen para componer los informes genéticos finales. Estos subprocesos comprenden fases relativamente complejas de integración de distinto software y tecnologías, así como de conocimiento de nomenclatura bioinformática que se encuentra en diversas bases de datos ómicas. Cada una de estas bases de datos aporta información diferente que es necesario analizar, contrastar y elaborar para una correcta presentación al usuario final.

1.2 Objetivos del Trabajo

En este apartado se van a definir tanto el objetivo principal del proyecto como un listado de objetivos secundarios que subyacen del objetivo principal.

1.2.1 Objetivos generales

El objetivo principal (OP) del proyecto es el diseño e implementación de una plataforma de generación de informes médicos personalizados en la que partiendo de datos SNPs se ofrezca información genética al usuario para que éste conozca las posibles enfermedades que puede padecer a lo largo de su vida, su origen o composición genética según datos geográficos, información nutricional, etc.

1.2.2 Objetivos secundarios

La consecución del objetivo general lleva asociada varios objetivos secundarios que se completan al desarrollar alguna de las distintas fases en las que se dividirá el proyecto.

- O1. Durante el trabajo a realizar en el proyecto será necesario entender conceptos ligados a anotaciones genéticas, cambios y mutaciones en genes. Por lo tanto, el conocimiento de estos conceptos es un objetivo secundario.
- O2. Por otra parte, también es objetivo secundario la búsqueda de información automatizada e integración entre las numerosas y diferentes bases de datos ómicas existentes.

- O3. Adicionalmente, es objetivo secundario la profundización en herramientas de desarrollo bioinformáticas y bioestadísticas, tales como R , el conocimiento de APIs y procedimientos de acceso a datos ómicos.
- O4. Estudio de “Linkage Disequilibrium” (LD). “Linkage Disequilibrium” es un fenómeno por el cual bloques del genoma se heredan de forma conjunta. Entre las diversas implicaciones que esto conlleva, es la variación de forma conjunta y la relación con una enfermedad de las variaciones de un mismo bloque genético, es decir, que se encuentren cerca físicamente. En este caso, se consideran todas una misma variante y se coge una en representación de todo el bloque.
- O5.Estudio de interacción entre genes para producir enfermedades humanas. En numerosas ocasiones las variantes (SNPs) no actúan de forma independiente, sino que interaccionan entre ellas para producir cambios nocivos.

Estos dos últimos objetivos secundarios (O4 y O5) se podrán llevar a cabo en función de la planificación final del proyecto.

1.3 Enfoque y método seguido

El objetivo es desarrollar un producto nuevo, a pesar de que ya existen algunas soluciones comerciales, el proyecto persigue el objetivo de realizar un prototipo que contemple todas las fases y pasos necesarios para la generación de informes genéticos. Además, el proyecto contiene un porcentaje alto de investigación e innovación, no únicamente trata de desarrollo de software.

Para abordar este proyecto, se propone seguir una metodología de proyecto dividida en fases, bien de desarrollo, bien de investigación. Cada una de esas fases tendrá unos resultados que servirán como “input” de las siguientes hasta la finalización del proyecto. Además, como el proyecto tiene una componente de investigación, hay que definir un plan de mitigación en caso de que los resultados de las investigaciones no sean lo suficientemente satisfactorios y haya que reconducir el proyecto en alguna fase del mismo.

1.4 Planificación del Trabajo

En los siguientes subapartados se explica el plan de trabajo de una forma detallada. Por una parte se definen las tareas principales, aportando además un coste temporal aproximado de cada una de ellas y un diagrama de planificación de las mismas.

Asimismo, se presentan los hitos marcados en el proyecto y los productos que se obtendrán en cada uno de ellos.

Por último se plantean estrategias de mitigación de riesgos y plan de contingencia para prevenir problemas durante la ejecución del proyecto.

1.4.1 Tareas

La siguiente tabla enumera las tareas principales que se realizarán en el proyecto. Cada tarea se relaciona con uno o varios objetivos marcados en los puntos anteriores del documento.

Identificador tarea	Tarea	Objetivos que cubre	Duración Aprox.
T1	Tarea 1. Estudio del estado del arte	Objetivo secundario 1. Comprensión del problema.	30 horas
T2	Tarea 2. Obtención de datos SNPs de prueba	Objetivos secundarios 2 y 3. Bases de datos y herramientas bioinformáticas.	40 horas
T3	Tarea 3. Análisis y selección de software para el desarrollo del proyecto	Objetivos secundarios 2 y 3. Bases de datos y herramientas bioinformáticas.	20 horas
T4	Tarea 4. Diseño de la arquitectura de la solución.	Objetivo principal.	24 horas
T5	Tarea 5. Desarrollo de módulo de extracción de información genética para enfermedades humanas	Objetivo principal	100 horas
T6	Tarea 6. Desarrollo de módulo de informes sobre ancestros	Objetivo principal	40 horas
T7	Tarea 7. Estudio de software para el análisis de “Linkage Disequilibrium”	Objetivo secundario 4. Estudio de “Linkage Disequilibrium” (LD)	20 horas
T8	Tarea 8. Desarrollo de módulo para detectar “Linkage Disequilibrium”	Objetivo secundario 4. Estudio de “Linkage Disequilibrium” (LD)	30 horas
T9	Tarea 9. Elaboración de memoria	Objetivo principal	40 horas

T10	Tarea 10. Elaboración de presentación	Objetivo principal	20 horas
Total			**404 horas**

Una vez listadas las tareas, dentro de la planificación del proyecto, es importante definir un alcance mínimo para considerarlo como satisfactorio, acotar la planificación global del proyecto y la estimación en esfuerzo de cada fase.

Se considera alcance mínimo la implementación del informe sobre enfermedades humanas partiendo de datos SNPs.

La siguiente tarea que aporta valor añadido al proyecto es el estudio de “Linkage Disequilibrium” la posibilidad de desarrollar un módulo que lo trate. Es por esto que esta tarea sería la primera a abordar después de conseguir el alcance mínimo en función del tiempo disponible.

Los informes adicionales de ancestros, nutrición, etc, aportan variedad a la plataforma e integración con más bases de datos y herramientas software pero no añaden ninguna componente de investigación, por lo tanto se estudiarán los procedimientos necesarios para su implementación y se desarrollarán en función del tiempo disponible. Esta tarea también es adicional al alcance mínimo del proyecto.

1.4.2 Calendario

A continuación se muestra un gráfico con la planificación dividida en tareas y semanas, marcando tanto los hitos como las fechas más importantes en cuanto a PECs según el plan docente.

Tarea	Descripción	Marzo		Abril			
		Semana 4	Semana 5	Semana 1	Semana 2	Semana 3	Semana 4
T1	Estudio del estado del arte	H1					
T2	Obtención de datos SNPs de prueba		H2				
T3	Análisis y selección de software para el desarrollo del proyecto			X			
T4	Diseño de la arquitectura de la solución			H3			
T5	Desarrollo de módulo de extracción de información genética para enfermedades humanas			X	X	X	H4 PEC2-23/04/18
T6	Desarrollo de módulo de informes sobre ancestros						
T7	Estudio de software para el análisis de "Linkage Disequilibrium"						
T8	Desarrollo de módulo para detectar "Linkage Disequilibrium"						
T9	Elaboración de memoria						
T10	Elaboración de presentación						

Tarea	Descripción	Mayo					Junio	
		Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 1	Semana 2
T1	Estudio del estado del arte							
T2	Obtención de datos SNPs de prueba							
T3	Análisis y selección de software para el desarrollo del proyecto							
T4	Diseño de la arquitectura de la solución							
T5	Desarrollo de módulo de extracción de información genética para enfermedades humanas							
T6	Desarrollo de módulo de informes sobre ancestros	X	H5					
T7	Estudio de software para el análisis de "Linkage Disequilibrium"			X				
T8	Desarrollo de módulo para detectar "Linkage Disequilibrium"			X	H6 PEC3-21/05/18			
T9	Elaboración de memoria					X	H7 PEC4-05/06/18	
T10	Elaboración de presentación							H8 PEC5-13/06/18

1.4.3 Hitos

La siguiente tabla refleja los hitos planificados para la correcta ejecución del proyecto así como su relación con las entregas especificadas en el plan docente de la asignatura.

Hito	Relación con PEC	Fecha de consecución
H0. Plan de trabajo	PEC 1	19/03/18
H1. Estudio del estado del arte		23/03/18
H2. Obtención de datos SNPs de prueba		30/03/18
H3. Diseño de la arquitectura de la solución		06/04/18
H4. Desarrollo de módulo de extracción de información genética para enfermedades humanas	PEC 2	23/04/18
H5. Desarrollo de módulo de informes sobre ancestros	PEC 3	11/05/18
H6. Desarrollo de módulo para detectar “Linkage Disequilibrium”		21/05/18
H7. Elaboración de memoria	PEC 4	05/06/18
H8. Elaboración de presentación	PEC 5	13/06/18

1.4.4 Análisis de riesgos

El proyecto consiste en desarrollar un producto nuevo, con un alto grado de investigación. Además el tiempo disponible para su desarrollo es relativamente corto, alrededor de 3 meses. Todos estos factores hacen que existan varios riesgos a tener en cuenta que pueden poner en peligro el éxito del mismo.

Es imperativo afinar mucho el alcance el proyecto, ya que un proyecto demasiado ambicioso bien por parte de investigación e innovación puede ser irrealizable o bien por la cantidad de informes a desarrollar o inalcanzable en tiempo.

Ante el riesgo de no conseguir obtener resultados en la parte innovación e investigación relativa al “Linkage Disequilibrium”, se debería suprimir la fase de implementación y dejar una sección teórica con la explicación del problema y algunas de las soluciones actuales si existieran.

En el caso de que el desarrollo de los informes tenga un sobreesfuerzo en horas muy superior a lo estimado inicialmente, se puede mitigar el riesgo de fracaso de tres formas principalmente:

1. Eliminando el número de informes a realizar. Por ejemplo suprimir el de nutrigenómica. Estos informes extra, siempre se pueden añadir como trabajo futuro.
2. Reduciendo las bases de datos con las que se integrará el generador de informes. Una vez finalizado el proyecto, se pueden añadir más bases de datos de integración como trabajo futuro.
3. Bajando la calidad de los informes, presentando menos información de la que inicialmente se podría esperar. El valor del producto es el conocimiento que aporta para el desarrollo de la pipeline de generación de informes, por lo tanto, de nuevo, la cantidad de datos a mostrar se podría ampliar como trabajo futuro.

1.5 Breve sumario de productos obtenidos

La tabla siguiente muestra el listado de hitos junto con el resultado esperado a la consecución de cada uno de ellos.

Hito	**Resultado**
H1. Estudio del estado del arte	Conocimiento de las empresas y servicios actuales de informes genéticos, así como de herramientas y "pipelines" bioinformáticas
H2. Obtención de datos SNPs de prueba	Conjunto de datos de SNPs para simular muestras humanas y poder realizar informes genéticos.
H3. Diseño de la arquitectura de la solución	Documento con los principales módulos que participan en el proyecto.
H4. Desarrollo de módulo de extracción de información genética para enfermedades humanas	Módulo software con funcionalidad para obtener datos relativos a enfermedades humanas partiendo de los datos de muestra.
H5. Desarrollo de módulos de informes extra. Ancestros y nutrigenética.	Módulo software con funcionalidad para obtener datos relativos a composición genética geográfica partiendo de los datos de muestra.También informes sobre nutrición.
H6. Desarrollo de módulo para detectar "Linkage Disequilibrium"	Módulo software que trata los casos de "Linkage disequilibrium"
H7. Elaboración de memoria	Memoria final del proyecto
H8. Elaboración de presentación	Presentación del proyecto

1.6 Breve descripción de los otros capítulos de la memoria

En el primer capítulo se presentará el estado del arte actual, los conceptos básicos y principales del mismo. Se corresponde con el hito 1.

La definición y análisis de las fuentes de datos se expondrá en el segundo capítulo de la memoria. Se seleccionará por un parte el conjunto de datos SNPs de prueba y por otra las bases de datos de las que se extraerá la información necesaria para los informes. Este capítulo cubre el hito 2.

El tercer capítulo relativo al hito 3 es un capítulo con contenido más técnico y abarca la definición de la arquitectura general de la solución y sus principales componentes.

El desarrollo, pruebas y validación de los informes, se corresponde con los hitos 4 y 5 del proyecto y se presenta en el capítulo cuarto.

El capítulo quinto, se reserva para la parte de investigación y desarrollo de "Linkage Disequilibrium" y de interacción entre genes para causar enfermedades humanas.La implementación de la capa de análisis de datos se expondrá en el capítulo quinto del documento.

Por último, el documento finalizará con un capítulo dedicado a conclusiones extraídas durante la ejecución del proyecto.

2. Resto de capítulos

2.1. Estado del arte

Conceptos generales

La genética humana describe el estudio de la herencia biológica en los seres humanos. La genética humana abarca gran variedad de campos como son la genética clásica, citogenética, genética molecular, biología molecular, genómica, genética de poblaciones, genética del desarrollo, genética médica y el asesoramiento genético. El estudio de la genética humana puede ser útil ya que puede responder preguntas acerca de la naturaleza humana, comprender el desarrollo eficaz para el tratamiento de enfermedades y la genética de la vida humana, entre otras muchas aplicaciones.

De los posibles campos enumerados anteriormente, el proyecto se centrará en la genética médica, asesoramiento genético y genética de poblaciones.

El ADN codifica el mapa genético completo de seres humanos. Es lo que hace a los miles de millones de personas en el planeta únicos, sin embargo, al mismo tiempo, genéticamente similares a los de sus padres y antepasados.

El ADN humano sufre variaciones genéticas. Las más comunes son las conocidas como “Single nucleotide polymorphisms” (SNPs). Cada SNP representa una diferencia en un único componente de la cadena de ADN, llamado nucleótido. Por ejemplo, un SNP puede reemplazar el nucleótido Citosina (C) por el nucleótido Timina (T) en una cierta posición de la cadena de ADN. Los estudios genéticos basados en variaciones genéticas se conocen como “Genome Wide Association Studies” (GWAS). GWAS típicamente se centran estudiar las asociaciones entre SNPs y rasgos y o enfermedades observadas en el fenotipo de humanos, aunque pueden realizarse sobre cualquier otro organismo.

Los SNPs ocurren, de media, una vez cada 300 nucleótidos, lo que significa que existen aproximadamente 10 millones de SNPs en el genoma humano. Estos SNPs actúan como biomarcadores, ayudando a los científicos a localizar genes que están asociados con enfermedades. Cuando ocurre un SNP dentro de un gen o cerca de una región regulatoria cercana a un gen, puede contribuir a la aparición o empeoramiento de una enfermedad afectando a la función del gen. Por lo tanto, el estudio de SNPs será la base del proyecto que se va a desarrollar.

En la parte de genealogía o genética de poblaciones, es importante mencionar los tipos de ADN que se utilizan para tales efectos y qué tipo de información proporciona cada uno. Son los siguientes:

- ADN de cromosoma (Y-DNA)
- El ADN mitocondrial (ADNmt)

- ADN autosómico

Y-ADN es un tipo de ADN que que analiza el **cromosoma Y**, que se hereda exclusivamente de padres a hijos, sigue únicamente la línea paterna y solo se lo pueden hacer los hombres.

Y-ADN es particularmente útil para el seguimiento de una línea directa paterna (padre, abuelo paterno, etc), ya que cambia lentamente de generación en generación.

ADNmt es un tipo de ADN que se hereda de madres a hijos e hijas y es abierto a ambos sexos. Esto hace ADNmt útil para el rastreo una de línea directa materna (madre, abuela materna, bisabuela materna, etc.)

ADN autosómico es el tipo de ADN responsable de las características más físicas, tales como altura, color de ojos, etc. El ADN autosómico
es heredado por los hijos y las hijas de ambos padres (y de sus cuatro abuelos, etc.).

Medicina genética como negocio

Dadas las posibilidades que ofrece la genética, y que los avances tecnológicos en el ámbito del análisis genético han reducido considerablemente tanto el coste económico como el temporal de un análisis de genoma humano, han ido apareciendo en estos últimos años varias empresas con líneas de negocio enfocadas a realizar análisis genéticos de diversas características. Este tipo de negocio es conocido también como test genético Direct-To-Consumer (DTC).

El interés que despierta en las personas conocer ciertos aspectos de su genética y su posible asociación al padecimiento de enfermedades, rendimiento deportivo, origen genético y tratamientos personalizados, entre otras, ha propiciado la popularidad de estos análisis.

En USA, ciertos informes como son los de ancestros, se han hecho bastante populares y sirven incluso como regalo para otras personas, es relativamente frecuente que una persona reciba por su cumpleaños un informe de ancestros para que conozca su composición y sus raíces genéticas. Algunas compañías a nivel mundial que realizan estos informes son 23andme (https://www.23andme.com), deCODE (https://www.decode.com) y FamilyTreeDNA (https://www.familytreedna.com).
En España también existen algunos ejemplos de compañías que generan negocio a partir de informes genéticos de varios tipos. Las compañías más conocidas y con negocio exclusivo en la generación de informes son Tellmegen (https://www.tellmegen.com) y 24Genetics (https://www.24genetics.com). Otra compañía que ofrece varios servicios genómicos y que incluyen varios informes genéticos es por ejemplo Sistemas Genómicos.

Software bioinformático

La evolución de la tecnología bioinformática y sobre todo de cómputo y cálculo distribuido en los últimos años, ha permitido que los análisis de ADN, secuenciación, anotación, etc, pasen de varios días de trabajo a algunas horas, dependiendo de las máquinas que se utilicen para ello. Es importante recalcar que el genoma humano consta de alrededor de 3.200 millones de bases, lo que traducido a almacenamiento informático supone unos 800 Mgs (equivalente a algo más de un CD). Por lo tanto cada análisis que se realiza tiene que evaluar y analizar esta cantidad de datos varias veces, porque los procesos bioinformáticos constan de varias fases.

En cuanto al software que se utiliza, actualmente las tecnologías con comunidades bioinformáticas especialmente activas y con evolución y desarrollo constante son Python, R y Perl. Estas tecnologías están basadas en lenguajes de scripting y por lo tanto interpretadas, en contraposición a otras tecnologías compiladas.

2.2. Análisis de fuentes de datos y herramientas software

Este capítulo tiene por objetivo exponer por una parte las alternativas disponibles para obtener datos, ya sean datos SNPs para pruebas como las diferentes bases de datos ómicas desde las que podemos obtener información relacionada con los SNPs a analizar, y por otra parte, estudiar y seleccionar las herramientas software que se utilizarán para el acceso a los datos ómicos.

2.2.1. Análisis de fuentes de datos

El primer objetivo es identificar una fuente de datos que permita obtener datos SNPs de prueba. El caso ideal es encontrar datos abiertos de personas que se hayan realizado una prueba de ADN.
El proyecto openSNP (https://opensnp.org) es plataforma en línea de crowdsourcing para clientes de DTC interesados en compartir sus datos brutos (raw data) y para investigadores interesados en realizar GWAS u otros tipos de análisis con los datos. Se alienta a los clientes de las pruebas de DTC a compartir sus resultados de genotipado junto con sus características fenotípicas para facilitar el acceso de los investigadores. Los usuarios de openSNP pueden crear un perfil personal, analizar SNP y fenotipos en la plataforma mediante un simple sistema de comentarios, o enviar mensajes privados.

Las personas interesadas en utilizar los datos de openSNP pueden descargar volcados completos de la información genotípica y fenotípica o utilizar su API de consulta utilizando objetos de JavaScript Object Notation (JSON) o el Sistema de Anotación Distribuida (DAS).

Actualmente, los usuarios pueden cargar sus resultados de genotipado de las empresas 23andMe, deCODEme y FamilyTreeDNA a través de una interfaz web al proyecto de openSNP. Los datos cargados se publican bajo la licencia

Creative Commons Zero, que, de acuerdo con los Principios de Panton (http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001195), permite una reutilización completa de los datos sin restricciones.

En openSNP, tal y como se explica en los párrafos anteriores, es posible subir datos de diversas empresas, cada empresa ofrece los resultados de SNPs con un formato diferente, aunque existen campos comunes y necesarios en los distintos tipos de archivo. Existe un estándar promovido por el proyecto "1000 Genomas" entre otros para la representación de variaciones genéticas conocido como Variant Call Format (VCF). En el momento de la redacción de este documento, VCF se encuentra en la versión 4.3. La especificación se puede encontrar en la url: http://samtools.github.io/hts-specs/VCFv4.3.pdf.

El segundo objetivo, consiste en seleccionar un subconjunto de bases de datos ómicas que puedan ofrecer información relevante a partir de los SNPs obtenidos en el punto anterior. Para realizar dicha selección hay que tener en cuenta el alcance del proyecto y el tiempo disponible para realizarlo. También será importante valorar la dificultad o facilidad de integración con otro software de estas fuentes de datos, ya que es un factor que afecta al tiempo de desarrollo.

El 'National Center for Biotechnology Information' (NCBI) proporciona una multi base de datos de biología molecular llamada ENTREZ. Esta base de datos conecta e integra otras bases de datos y recursos para poder relacionar la información proporcionada por cada una de las bases de datos de forma individual.

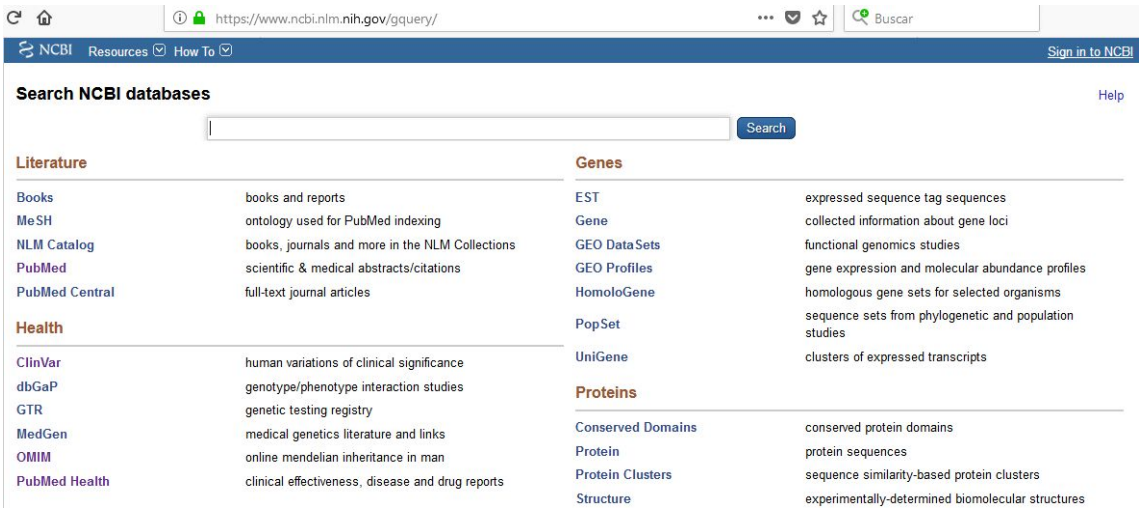
En el momento de elaboración de este documento ENTREZ constaba de 67 bases de datos que pueden explorarse a partir de la url https://www.ncbi.nlm.nih.gov/guide/all.

Para la realización de este proyecto, se han identificado como bases de datos indispensables las siguientes:

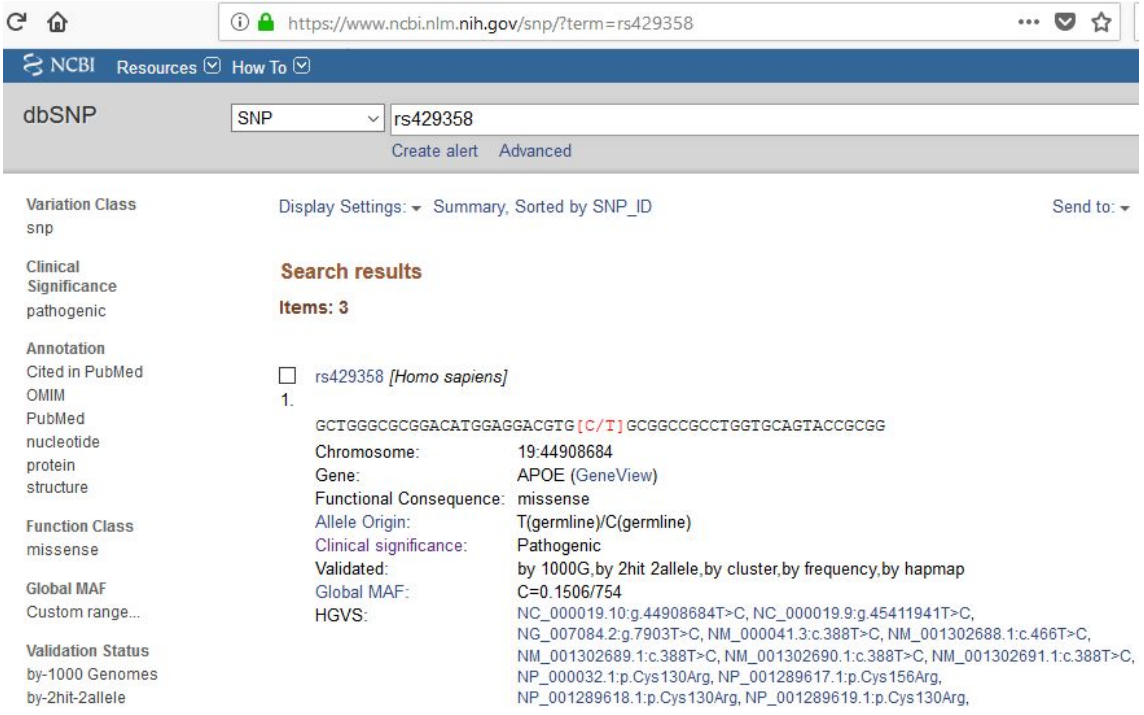
Base de datos	**Propósito (en inglés)**
Database of Short Genetic Variations (dbSNP)	Includes single nucleotide variations, microsatellites, and small-scale insertions and deletions. dbSNP contains population-specific frequency and genotype data, experimental conditions, molecular context, and mapping information for both neutral variations and clinical mutations.
ClinVar	A resource to provide a public, tracked record of reported

	relationships between human variation and observed health status with supporting evidence. Related information in the NIH Genetic Testing Registry (GTR), MedGen, Gene, OMIM, PubMed and other sources is accessible through hyperlinks on the records.
Database of Genomic Structural Variation (dbVar)	The dbVar database has been developed to archive information associated with large scale genomic variation, including large insertions, deletions, translocations and inversions. In addition to archiving variation discovery, dbVar also stores associations of defined variants with phenotype information.
Database of Genotypes and Phenotypes (dbGaP)	An archive and distribution center for the description and results of studies which investigate the interaction of genotype and phenotype. These studies include genome-wide association (GWAS), medical resequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits.
Genes and Disease	Summaries of information for selected genetic disorders with discussions of the underlying mutation(s) and clinical features, as well as links to related databases and organizations.
Online Mendelian Inheritance in Man (OMIM)	A database of human genes and genetic disorders. NCBI maintains current content and continues to support its searching and integration with other NCBI databases. However, OMIM now has a new home at omim.org, and users are directed to this site for full record displays.

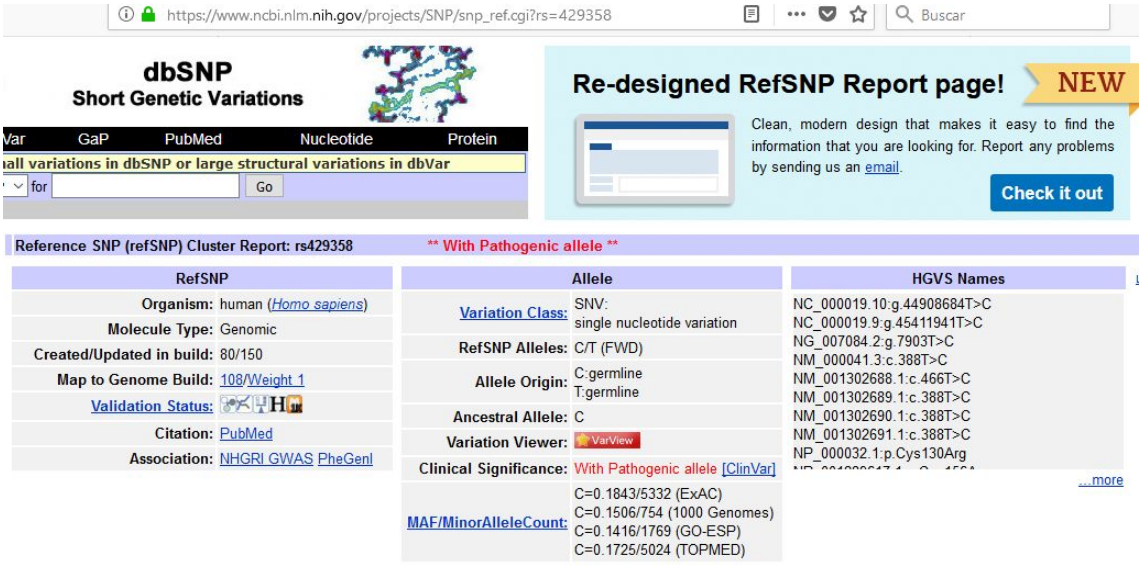
La imagen siguiente muestra la pantalla de inicio de la web de búsqueda en Entrez. Tal y como puede apreciarse existen numerosas bases de datos, desde relacionadas con publicaciones hasta de salud, genes,proteínas, etc.



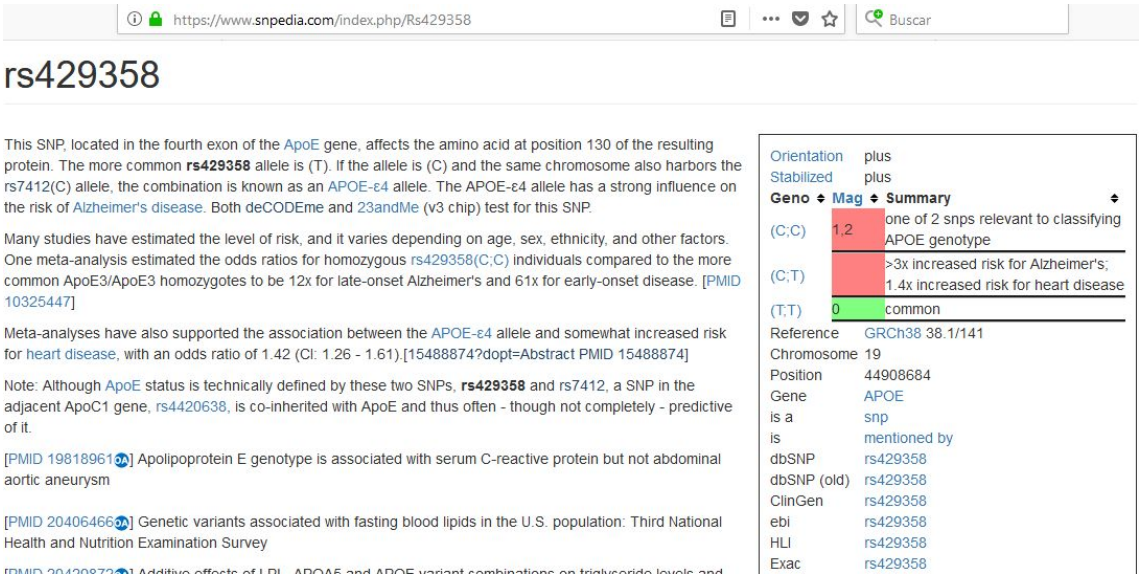
Al realizar una búsqueda sobre la base de datos SNP el término rs429358, obtenemos 3 resultados (sólo vemos uno en la imagen siguiente):



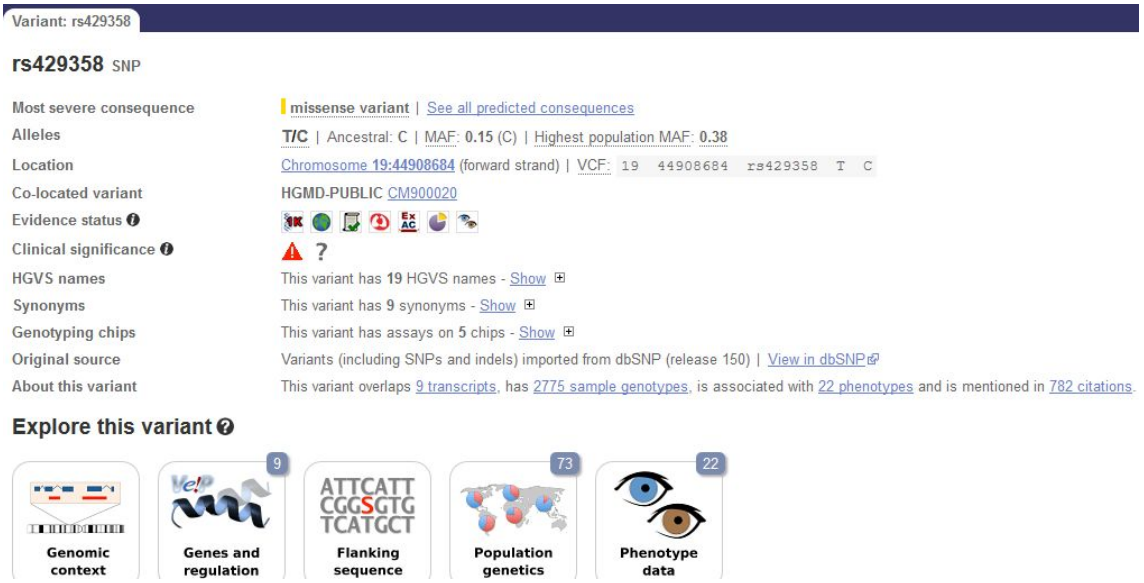
Y si seguimos el enlace al primer resultado podemos ver que existe información sobre significado cínico que indica que puede ser una variante patógena:



Otra base de datos interesante y que aglutina y enlaza diversa información a partir de un SNP es SNPedia (https://www.snpedia.com/index.php/SNPedia). La integración con esta base de datos queda supeditada al tiempo de desarrollo que cueste y a la disponibilidad y utilidad de información ofrecida que no se pueda conseguir a través de Entrez.
A continuación se adjunta un pantallazo con la búsqueda del SNP identificado como rs429358.



Por último, también se ha explorado la base de datos “Ensembl”, que proporciona información muy similar a las anteriores, tal y como mostramos en la imagen siguiente:



Para finalizar este apartado, se va a seleccionar la base de datos ENTREZ del NCBI como base de datos base o principal a la hora de desarrollar el proyecto. Esta selección se ha realizado tomando como criterios los siguientes:

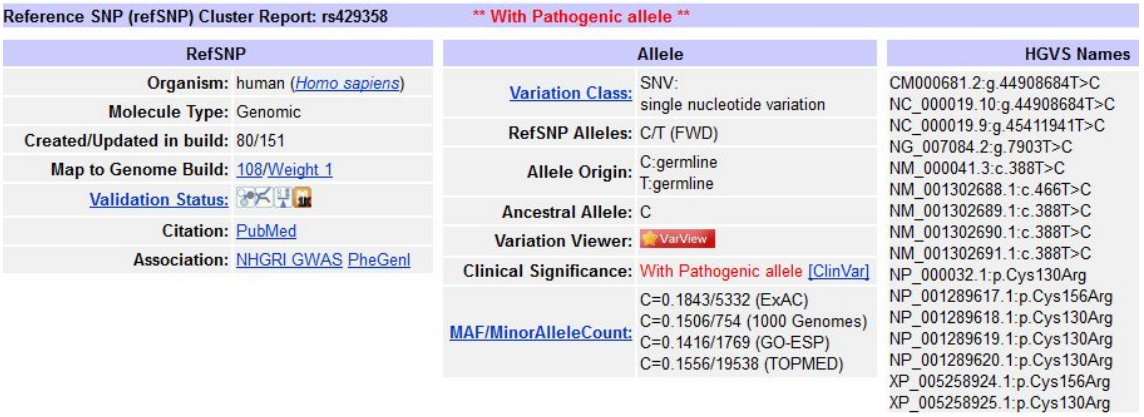
- Número de bases de datos disponibles.
- Frecuencia de actualización.
- Organismo que lo gestiona.
- Soporte y documentación disponible.

2.2.2. Exploración de datos disponibles para la realización de informes.

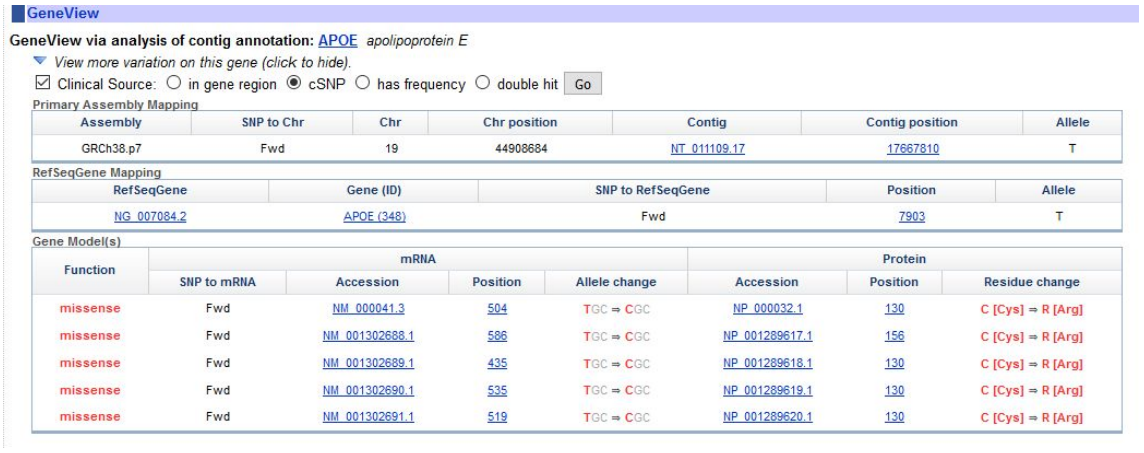
En el alcance del proyecto se citan algunas de las posibilidades que se pueden ofrecer a partir de la información asociada a un SNP. Es necesario, por lo tanto, analizar los datos disponibles al realizar una consulta partiendo de un SNP para poder acotar los tipos de informes que se realizarán.

Desde la base de datos SNP, se ofrece la siguiente información, relacionada la mayor parte de ella con enlaces a otras bases de datos.

- Reference SNP (refSNP) Cluster Report. Presenta datos generales del SNP.



- GeneView. Ofrece detalles sobre el Gen, posición de la mutación, nucleótidos en la cadena, etc.



GeneView

GeneView via analysis of contig annotation: APOE *apolipoprotein E*

View more variation on this gene (click to hide).

Clinical Source: in gene region | cSNP | has frequency | double hit | Go

Primary Assembly Mapping

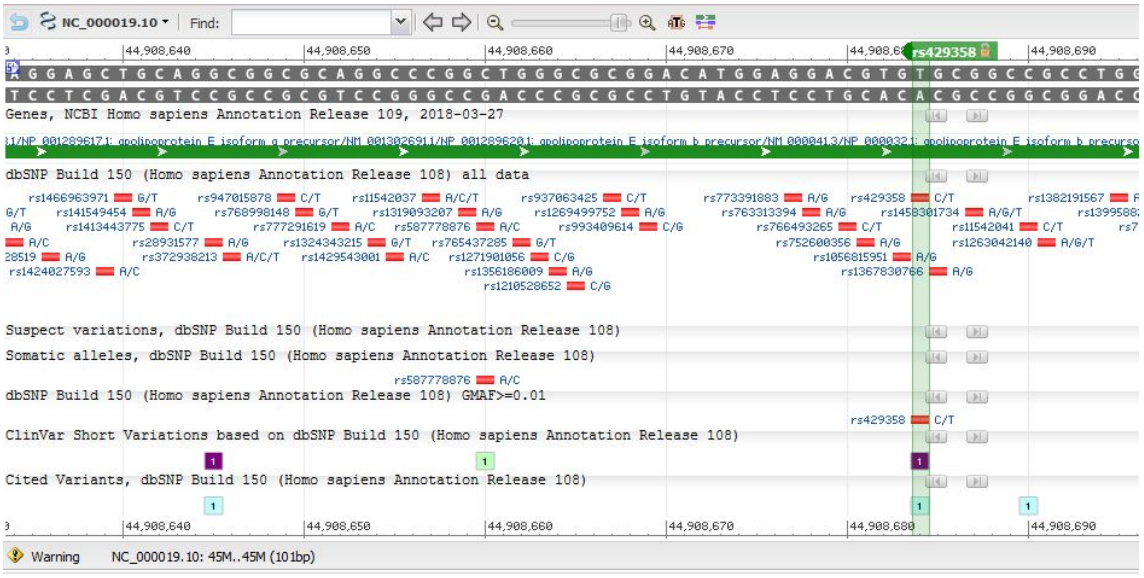
Assembly	SNP to Chr	Chr	Chr position	Contig	Contig position	Allele
GRCh38.p7	Fwd	19	44908684	NT_011109.17	17667810	T

RefSeqGene Mapping

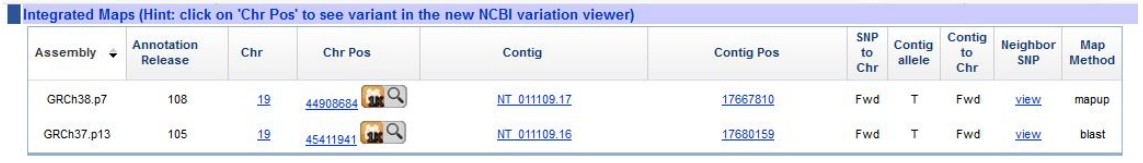
RefSeqGene	Gene (ID)	SNP to RefSeqGene	Position	Allele
NG_007084.2	APOE (348)	Fwd	7903	T

Gene Model(s)

Function	mRNA				Protein		
	SNP to mRNA	Accession	Position	Allele change	Accession	Position	Residue change
missense	Fwd	NM_000041.3	504	TGC ⇒ CGC	NP_000032.1	130	C [Cys] ⇒ R [Arg]
missense	Fwd	NM_001302688.1	586	TGC ⇒ CGC	NP_001289617.1	156	C [Cys] ⇒ R [Arg]
missense	Fwd	NM_001302689.1	435	TGC ⇒ CGC	NP_001289618.1	130	C [Cys] ⇒ R [Arg]
missense	Fwd	NM_001302690.1	535	TGC ⇒ CGC	NP_001289619.1	130	C [Cys] ⇒ R [Arg]
missense	Fwd	NM_001302691.1	519	TGC ⇒ CGC	NP_001289620.1	130	C [Cys] ⇒ R [Arg]

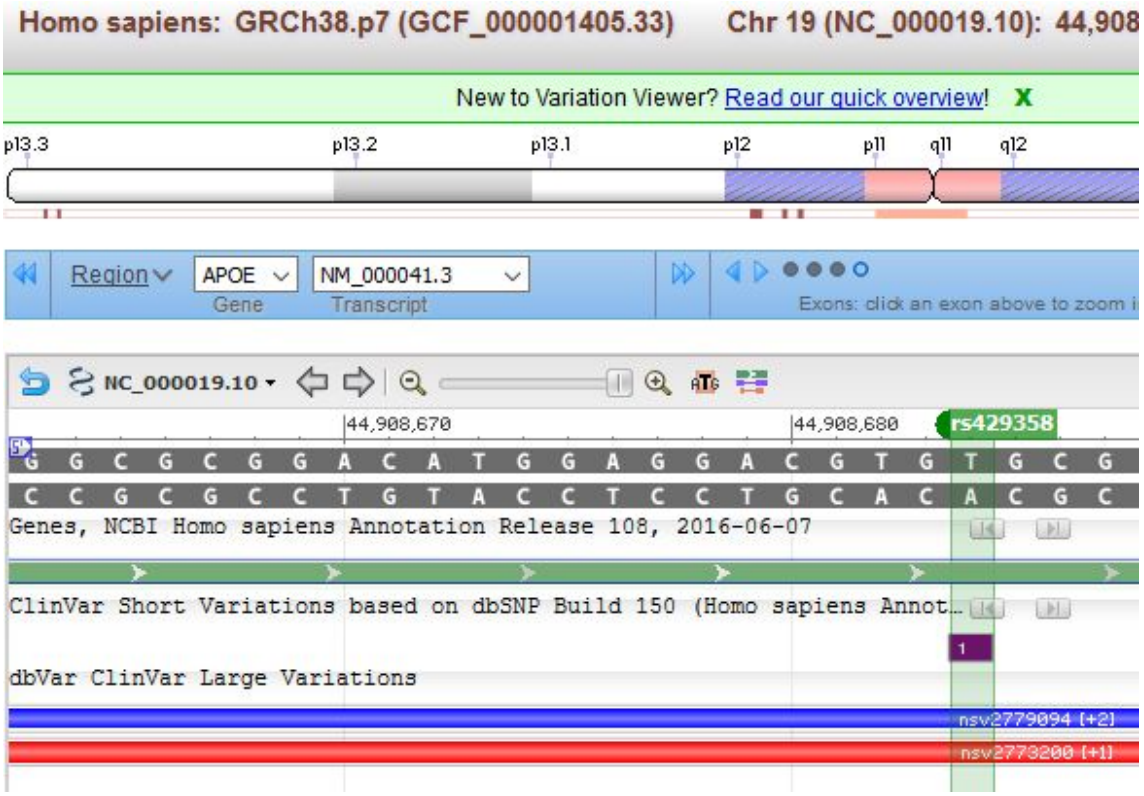


- Map. Muestra datos sobre la variación (posición, dirección del 'strand' o cadena de ADN, alelo, etc), así como pulsando en el enlace 'ChrPos', nos lleva a un visor de la cadena de ADN donde se encuentra la variación.



Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh38.p7	108	19	44908684	NT_011109.17	17667810	Fwd	T	Fwd	view	mapup
GRCh37.p13	105	19	45411941	NT_011109.16	17680159	Fwd	T	Fwd	view	blast



- Submission. Datos sobre las muestras de esta variación.

Submitter records for this RefSNP Cluster

The submission **ss76884559** has the longest flanking sequence of all cluster members and was used to instantiate sequence for **rs429358** during BLAST analysis for the current build.

NCBI Assay ID	Handle|Submitter ID	Validation Status	ss to rs Orientation /Strand	Alleles	5' Near Seq 30 bp	3' Near Seq 30 b
ss569295	SC_JCM|AC011481.2_65149		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss803061	SC_JCM|AF050154.1_21250		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss870163	DEBNICK|ae3937		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss2419938	HGBASE|SNP000002328		fwd/B	C/T	gctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss12568607	CUORCGL|SNP1		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss16231123	CGAP-GAI|1470380		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss21518782	SSAHASNP|WGSA-200403-chr19.chr19.NT_011109.15_17680159		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss24811489	SEQUENOM|sqnm198707		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss44158325	ABI|hCV3084793		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss76884559	SI_EXO|NT_011109.15_17680159		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss80743998	KRIBB_YJKIM|KHS1001043		fwd/B	C/T	gccmggctgggcgcggacatggaggacgtg	gcggcmgcctggtgcagtaccr
ss96308980	HUMANGENOME_JCVI|1103691153316		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss107936537	RSG_UW|APOE-004874		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg
ss132769779	ENSEMBL|ENSSNP1536174		fwd/B	C/T	gcccggctgggcgcggacatggaggacgtg	gcggccgcctggtgcagtaccg

- Fasta. Representación de la variación en formato FASTA.

Fasta sequence (Legend)

```
>gnl|dbSNP|rs429358|allelePos=401|totalLen=801|taxid=9606|snpclass=1|alleles='C/T'|mol=Genomic|build=151

CCTCGGCCTC CCAAAGTGCT GGGATTAGAG GCATGAGCCA CCTTGCCCGG CCTCCTAGCT
CCTTCTTCGT CTCTGCCTCT GCCCTCTGCA TCTGCTCTCT GCATCTGTCT CTGTCTCCTT
CTCTCGGCCT CTGCCCCGTT CCTTCTCTCC CTCTTGGGTC TCTCTGGCTC ATCCCCATCT
CGCCCGCCCC ATCCCAGCCC TTCTCCCCGC CTCCCACTGT GCGACACCCT CCCGCCCTCT
CGGCCGCAGG GCGCTGATGG ACGAGACCAT GAAGGAGTTG AAGGCCTACA AATCGGAACT
GGAGGAACAA CTGACCCCGG TGGCGGAGGA GACGCGGGCA CGGCTGTCCA AGGAGCTGCA
GGCGGCGCAG GCCCGGCTGG GCGCGGACAT GGAGGACGTG
Y
GCGGCCGCCT GGTGCAGTAC CGCGGCGAGG TGCAGGCCAT GCTCGGCCAG AGCACCGAGG
AGCTGCGGGT GCGCCTCGCC TCCCACCTGC GCAAGCTGCG TAAGCGGCTC CTCCGCGATG
CCGATGACCT GCAGAAGCGC CTGGCAGTGT ACCAGGCCGG GGCCCGCGAG GGCGCCGAGC
GCGGCCTCAG CGCCATCCGC GAGCGCCTGG GGCCCCTGGT GGAACAGGGC CGCGTGCGGG
CCGCCACTGT GGGCTCCCTG GCCGGCCAGC CGCTACAGGA GCGGGCCCAG GCCTGGGGCG
AGCGGCTGCG CGCGCGGATG GAGGAGATGG GCAGCCGGAC CCGCGACCGC CTGGACGAGG
TGAAGGAGCA GGTGGCGGAG GTGCGCGCCA AGCTGGAGGA
```

- Resource. Enlaces a otros recursos dentro de NCBI. También ofrece un resumen y un enlace a la base de datos 'ALFRED', donde se puede encontrar más información sobre la diversidad poblacional.

NCBI Resource Links

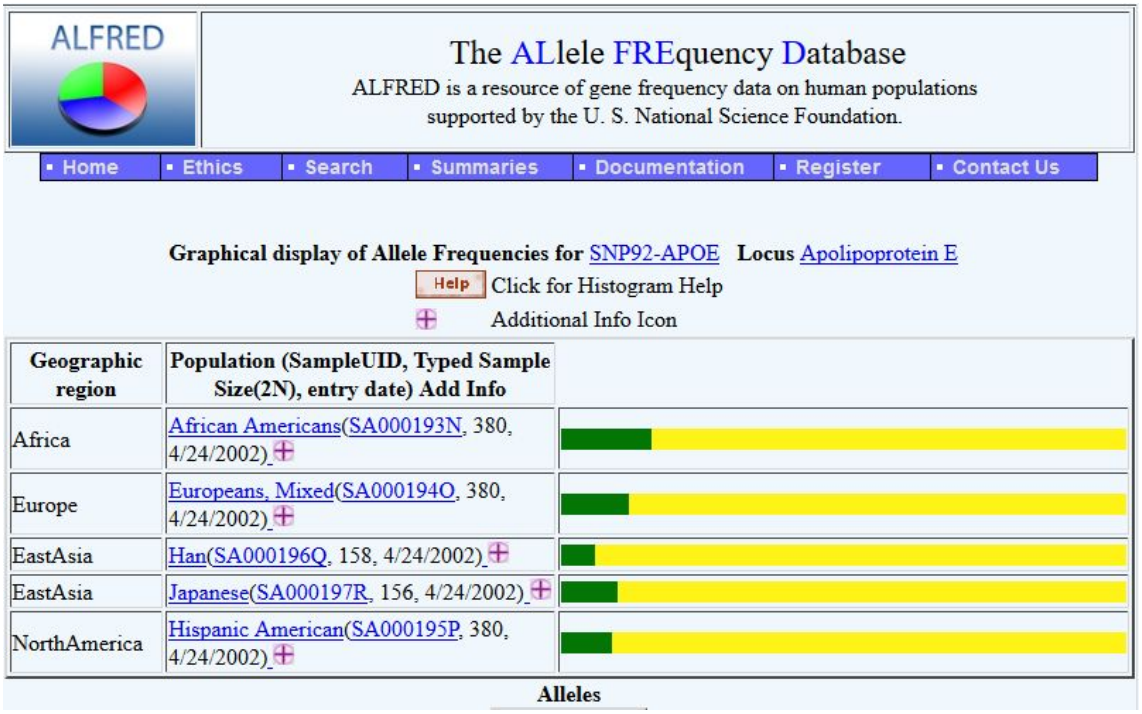
Submitter-Referenced		dbSNP Blast Analysis	UniGene Cluster ID	3D structure mapping	OMIM
dbSTS	GenBank				107741.0008
sqnm198707	NT_011109.15 ABBA01040892 AC011481 AC021988 BQ712095		515465	NP_000032	107741.0015
					107741.0016

- Diversity. Muestra una tabla con la diversidad poblacional de la variación.

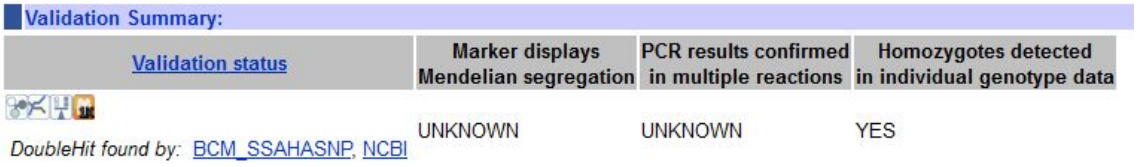
Population Diversity (Alleles in RefSNP orientation) . See additional population frequency from 1000Genome [here]

	Sample Ascertainment				Genotype Detail				Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	C/C	C/T	T/T	HWP	C	T
ss107936537	ABECASIS_CLINICAL_PANEL		752	AF					0.10638298	0.89361703
ss12568607	American Caucasians		1858	AF					0.14899999	0.85100001
ss132769779	ENSEMBL_Venter		2	IG		1.00000000			0.50000000	0.50000000
ss1363326184	EAS		1008	AF					0.08630000	0.91369998
	EUR		1006	AF					0.15509999	0.84490001
	AFR		1322	AF					0.26780000	0.73219997
	AMR		694	AF					0.10370000	0.89630002
	SAS		978	AF					0.08690000	0.91310000
ss168243995	CEU	European	2	IG		1.00000000			0.50000000	0.50000000

Summary	Average Het.+/- std err:	Individual Count	Founders Count	Individual Overlap	Genotype Conflict	Additional Freq. Data
	0.301+/-0.245	0	0	0	0	ALFRED: The Allele Frequency Database



- Validation. Información rellativa a la validación realizada sobre la variación.



Analizando esta información, los informes que se van a realizar en una primera versión de la plataforma son los de enfermedades humanas y los de ancestros, a través de la información que se presenta en la sección de diversidad poblacional.

2.2.3. Estudio y selección de software de acceso a datos

Una vez seleccionada la base de datos que se va a utilizar, el siguiente punto de estudio es la selección de herramientas y posibilidades de conexión para extraer la información de una forma automatizada de la base de datos.

En relación al acceso a 'Entrez', NCBI ha creado varios puntos de acceso llamados 'E-Utilities' y además proporciona información sobre estas herramientas en varios formatos. Por una parte, existe un recurso web (https://www.ncbi.nlm.nih.gov/books/NBK25501), donde se expone el manual de uso, mientras que adicionalmente NCBI ha generado videotutoriales en la plataforma youtube sobre 'E-Utilities'. Los enlaces a los videotutoriales se encuentran detallados en la sección de bibliografía.

Para facilitar la utilización de estas utilidades proporcionadas por el NCBI desde el lenguaje de programación R, se han encontrado algunos paquetes que ofrecen unos métodos de acceso más sencillos para extraer la información solicitada a Entrez. Entre los paquetes R analizados destacan:

- rentrez (https://cran.r-project.org/web/packages/rentrez/vignettes/rentrez_tutorial.html)
- reutils (https://cran.r-project.org/web/packages/reutils/reutils.pdf)

Opcionalmente y condicionado por la integración final o no de SNPedia, se ha encontrado un paquete de Bioconductor capaz de extraer información de SNPedia.

- Bioconductor (https://www.bioconductor.org)
- Bioconductor-SNPediaR (https://www.bioconductor.org/packages/release/bioc/html/SNPediaR.html)

Otro software que se ha visto y analizado siendo finalmente descartado ha sido:

- Phegeeni (https://www.youtube.com/watch?v=v_yEy--HcKc)

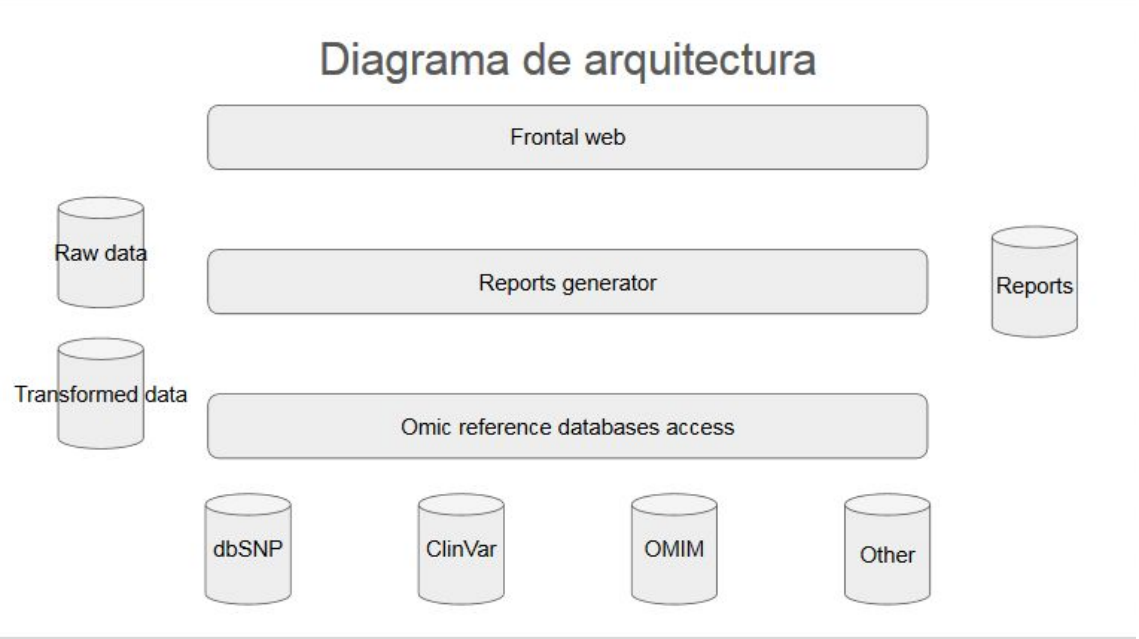
2.3. Arquitectura de la solución

En esta sección se presentan el diseño y la arquitectura a alto nivel de la solución. Por un lado se definirán los componentes principales y por otro se proporcionará un esquema con la cadena de trabajos (pipeline) para conseguir el resultado final. La última sección del capítulo resume las tecnologías utilizadas para el desarrollo de cada uno de los componentes de la solución.

El gráfico mostrado a continuación refleja los diferentes componentes que conforman en su conjunto el proyecto.

2.3.1. Arquitectura

La imagen de a continuación muestra los componentes principales de la arquitectura del proyecto.



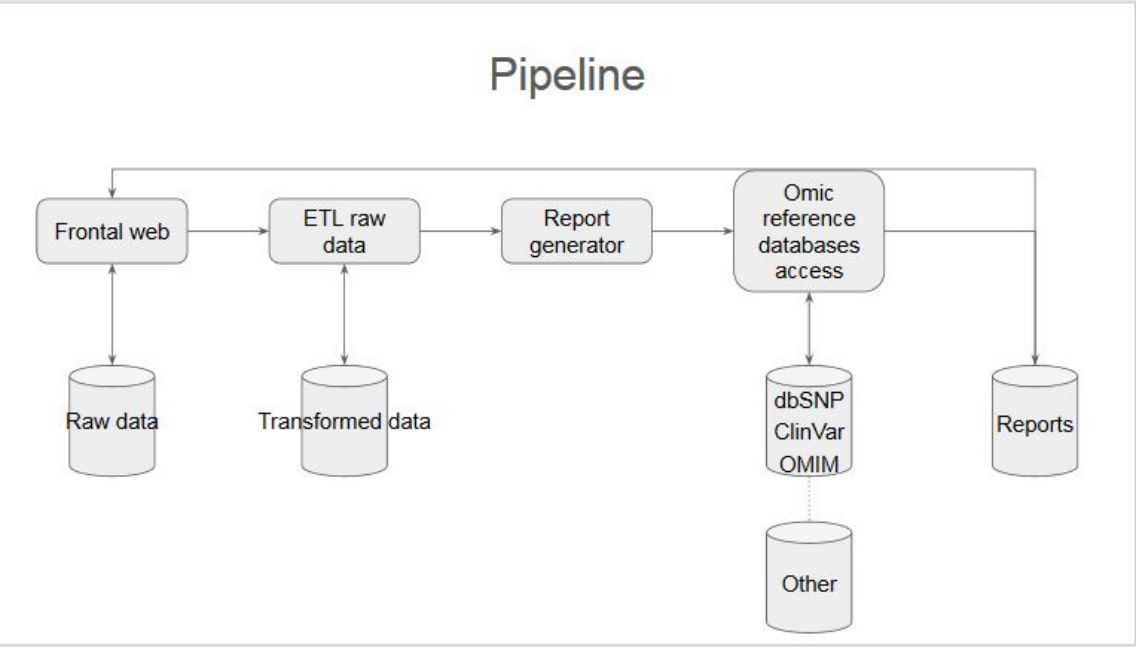
La arquitectura expuesta anteriormente consta de:

Componente	Función
Frontal web	Aplicación web que permitirá subir ficheros con 'raw data' y/o introducir SNPs en el formato especificado en el proyecto para la generación informes.
Report generator	Modulo principal para generar informes.
Omic reference databases access	Submódulo del generador de informes que obtiene información de

	las bases de datos ómicas de referencia utilizadas en el proyecto.
Repositorio de ‘raw data’	Mantiene los ficheros originales que se proporcionan desde el frontal web.
Repositorio de datos transformados	Guardará los datos transformados a partir de los ‘raw data’.
Repositorio de informes	Almacena los informes generados.

2.3.2. Workflow y Pipeline

En la siguiente ilustración puede verse el funcionamiento de la aplicación.
El proceso comienza en el frontal web, pasa una serie de etapas y finalmente la información generada como resultado vuelve a mostrarse desde la web.



Introducción y validación de datos

La primera etapa es la de introducción de datos, que se llevará a cabo bien especificando un fichero de ‘raw data’, bien introdución los SNPs de forma manual. Ambas opciones deben cumplir con las condiciones de formato soportadas por la aplicación y que se explicarán en los próximos párrafos.
Al realizar una exploración manual de la base de datos openSNP que contiene los ficheros de “raw data”, se ha visto que existen ficheros con diferente formato, y que no siguen el estándar VCF de representación de SNPs mencionado en capítulos anteriores, pero sí que constan de varios campos comunes que se pueden utilizar para unificar un formato concreto para esta solución. Los campos comunes encontrados en la exploración manual de los

ficheros y que utilizaremos durante el desarrollo de la solución son los siguientes:

Cabecera	Significado	Valor
fileId	Identificador del fichero de raw data desde el que se ha generado este fichero.	Cadena de caracteres
rsId	Identificador de SNP, si existe. Por lo tanto puede que venga vacío y se utilizará para comprobar el resultado de búsqueda en las bases de datos ómicas.	Cadena de caracteres (su valor puede venir especificado con un guión '-' si se desconoce)
chr	Cromosoma en el que se encuentra la variación genética.	Numérico
position	La posición de la variación genética en la cadena de ADN del cromosoma.	Numérico
allele 1	El alelo 1	Caracter (A, C, T, G, -)
allele 2	El alelo 2	Caracter (A, C, T, G, -)

Por lo tanto, el formato mínimo de una línea que defina un SNP debe tener la columna fileId y rsId de tipo cadena de caracteres, la columna 'chr' y la columna 'position', estos dos último de tipo numérico.

Algunos ejemplos de ficheros válidos extraídos de openSNP durante la inspección manual de la base de datos son:

Fichero 7211.ftdna-illumina.5593

 # MyHeritage DNA raw data.
 # This file was generated on 2017-09-20 15:17:33
 # For each SNP, we provide the identifier, chromosome number, base pair position and genotype.The genotype is reported on the forward (+) strand with respect to the human reference build 37.
RSID,CHROMOSOME,POSITION,RESULT
"rs4477212","1","82154","AA"
"rs3094315","1","752566","--"
"rs3131972","1","752721","AG"
"rs12562034","1","768448","--"
"rs12124819","1","776546","--"

Fichero 7208.ancestry.5590

#AncestryDNA raw data download
#This file was generated by AncestryDNA at: 12/19/2017 01:02:23 UTC
#Data was collected using AncestryDNA array version: V2.0
#Data is formatted using AncestryDNA converter version: V1.0
#Genetic data is provided below as five TAB delimited columns. Each line
#corresponds to a SNP. Column one provides the SNP identifier (rsID where
#possible). Columns two and three contain the chromosome and basepair position
#of the SNP using human reference build 37.1 coordinates. Columns four and five
#contain the two alleles observed at this SNP (genotype). The genotype is reported
#on the forward (+) strand with respect to the human reference.
rsid chromosome position allele1 allele2
rs190214723 1 693625 T T
rs3131972 1 752721 A G
rs12562034 1 768448 G G

Fichero 7210.23andme.5592

 # This data file generated by 23andMe at: Fri Mar 16 21:11:56 2018
#
 # This file contains raw genotype data, including data that is not used in 23andMe reports.
 # Below is a text version of your data. Fields are TAB-separated
 # Each line corresponds to a single SNP. For each SNP, we provide its identifier
 # (an rsid or an internal id), its location on the reference human genome, and the
 # genotype call oriented with respect to the plus strand on the human reference sequence.
 # We are using reference human assembly build 37 (also known as Annotation Release 104).
 # Note that it is possible that data downloaded at different times may be different due to ongoing
 # improvements in our ability to call genotypes. More information about these changes can be found at:
 # https://you.23andme.com/p/4b6e8480a2c8ff99/tools/data/download/
#
 # More information on reference human assembly build 37 (aka Annotation Release 104):
 # http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606
#
 # rsid chromosome position genotype
rs548049170 1 69869 TT
rs13328684 1 74792 --
rs9283150 1 565508 AA
i713426 1 726912 --

Una vez validado el formato de los datos introducidos, éstos se almacenan en un repositorio de ficheros para un procesado posterior.

Transformación de datos

Debido a la heterogeneidad de los ficheros existentes en openSNP, se incluye una tarea que va a generar datos homogéneos para nuestra aplicación, transformando los ficheros almacenados en 'raw data' y guardándolos en otro almacén denominado 'datos transformados'. A los procesos de transformación de información se les conoce como procesos ETL (Extract, Transform and Load). Los ficheros del almacén 'datos transformados' estarán formateados en concordancia con el formato definido para la aplicación:

#rsId #chr #position #genotype
rs000 1 1 TT

Análisis de datos

Esta subtarea tiene como entrada de datos los ficheros almacenados en el almacén 'datos transformados' y los procesa línea a línea extrayendo el identificador del cromosoma y la posición de la variación genética para consultar estos dos valores en las bases de datos ómicas de referencia.
Tras la consulta, se analizan los resultados y se guardan en el almacén de 'informes' en formato .csv para facilitar su posterior tratamiento en la visualización del informe.

Generación de informes

Tomando como entrada los resultados obtenidos en la fase de análisis de datos, y leyendo los ficheros csv del almacén de 'informes', la subtarea de generación de informes generará el informe asociado al fichero original de entrada y lo mostrará en el frontal web.

2.3.3. Tecnologías para el desarrollo

A continuación se ofrece un listado con las diferentes tecnologías utilizadas para el desarrollo del producto. Como puede apreciarse, se han integrado y orquestado varias tecnologías diferentes, aprovechando las ventajas de cada una de ellas para ofrecer una solución robusta y eficiente.

Frontend

- Lenguaje de programación Javascript, como tecnología base de desarrollo.
- Framework de desarrollo NodeJS y Express, sirven como base para crear un servidor web.
- Lenguaje de programación HTML y EJS para la generación de webs dinámicas mediante javascript

Proceso ETL

- Lenguaje de programación R como base de los procesos ETL.
- Transformaciones de datos y dataframes, lecturas y escrituras de ficheros, tratamiento de cadenas, utilizando varios paquetes del núcleo de R.
- Software PLINK para transformación entre formatos de datos.

Acceso a bases de datos ómicas

- Lenguaje de programación R como base de este módulo.
- Paquetes R como 'rentrez', 'reutils', 'snpedia', 'ensembl', para el acceso a bases de datos ómicas y extracción de información relevante.

Generador de informes

- Lenguaje de programación R para generar los ficheros definitivos con los informes.
- Paquete NodeJS para lanzar procesos RScript desde el servidor y recueprar la salida del proceso RScript.

2.4. Desarrollo de la solución

En este apartado se van a desarrollar los módulos que forman la plataforma de informes. Cada uno de los subapartados describirá el proceso seguido y las decisiones de implementación realizadas en cada módulo.

Limitaciones de implementación.

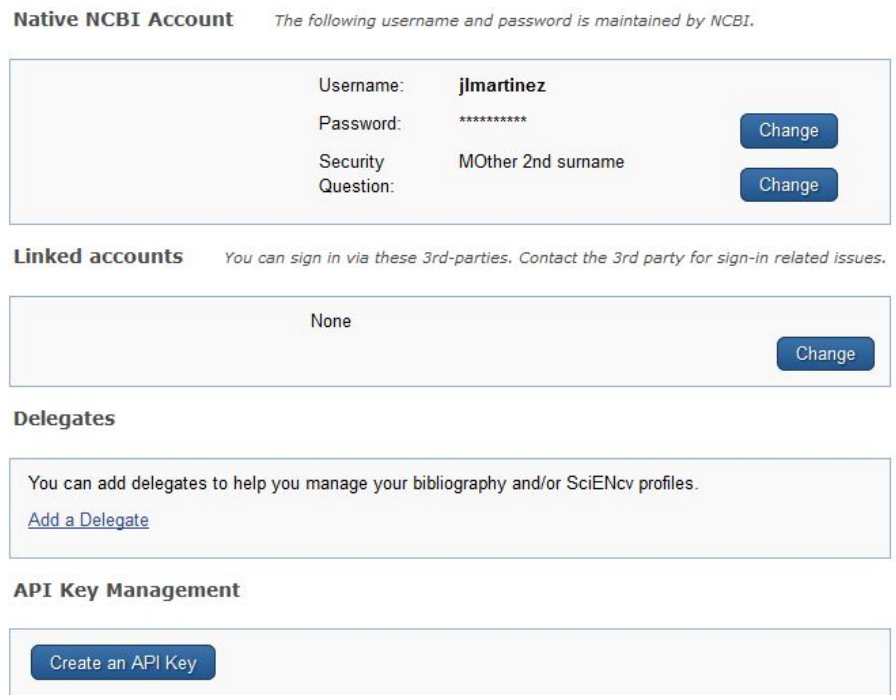
La base de datos Entrez de NCBI limita el acceso a aplicaciones no registradas y puede bloquear las peticiones si pasan de un límite de 3 peticiones por segundo. En nuestro caso, al procesar automáticamente varios SNP y realizar para cada SNP varias llamadas, se superará seguramente este ratio, por lo que a partir de Mayo de 2018, será necesario registrar la aplicación y conseguir una API key para utilizar Entrez. Una vez registrada la aplicación y conseguida la API Key se podrá llegar a un ratio de petición de 10 peticiones por segundo.
Estos límites y explicaciones sobre Entrez vienen detallados en los siguientes enlaces web.

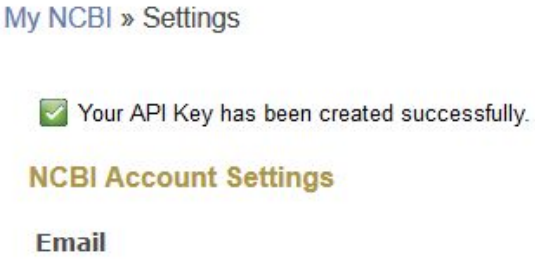
https://www.ncbi.nlm.nih.gov/books/NBK25497/
https://support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/2039/Why-and-how-should-I-get-an-API-key-to-use-the-E-utilities
https://www.r-bloggers.com/a-rentrez-paper-and-how-to-use-the-ncbis-new-api-keys/

Por lo tanto, en el código fuente, como medida de prevención, entre una llamada y otra se va a establecer un retraso de 0,3 segundos, para no sobrepasar las 3 peticiones por segundo permitidas.

A continuación muestro el proceso de registro en NCBI, para prevenir problemas en las fases de desarrollo de la plataforma.

Una vez registrado, desde la página de settings en la cuenta creada, hay que generar una API key:





Un escenario similar se presenta con el acceso a Ensembl, aunque las peticiones simultáneas permitidas no están especificadas a priori, o no se ha encontrado documentación relativa a este aspecto. El bloqueo de peticiones se informa al usuario mediante cabeceras HTTP, por lo que hay que analizar esa respuesta y esperar el tiempo que nos marcan en las cabeceras para continuar con el funcionamiento del programa.

https://github.com/Ensembl/ensembl-rest/wiki/Rate-Limits

2.4.1. Desarrollo del front end

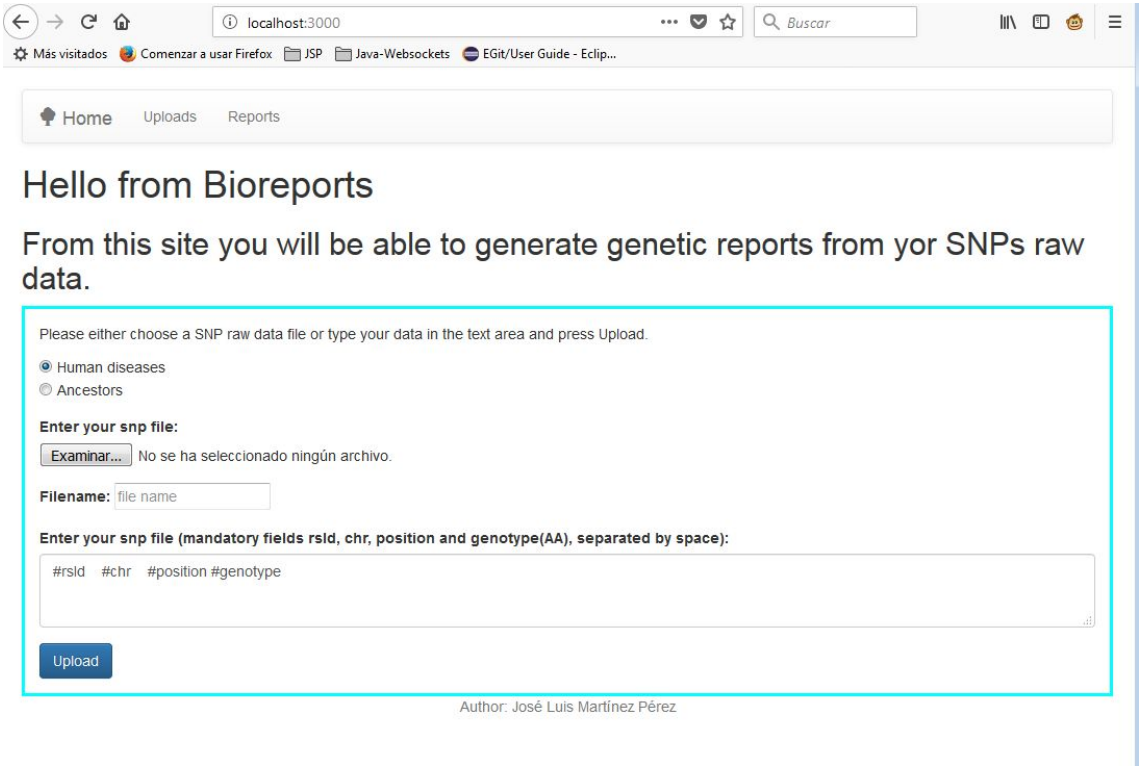
Para que los usuarios de la plataforma puedan aportar sus datos SPNs es necesario proporcionar una aplicación sencilla y usable.
Esta aplicación deberá aportar las siguientes funcionalidades, útiles para los usuarios y los administradores:

- Formulario de adquisición de datos.
    - Los datos SNP se podrán subir vía fichero o bien escribiendo manualmente los datos en un formulario. Si se proporcionan ambos datos, el fichero tendrá prioridad sobre los datos escritos en la caja de texto.
- Listado de ficheros subidos.
    - Sirve para que los administradores de la plataforma puedan ver los ficheros que se han subido a la aplicación.
- Listado de reportes listos para mostrar.
    - Los usuarios podrán los informes generados.
- Detalle de informe.
    - Los usuarios podrán ver los datos del informe seleccionado del listado de informes disponibles.

Formulario de adquisición de datos.

Es la pantalla principal de adquisición de fuentes de datos. A través de ella, el usuario proporcionará bien mediante un fichero existente o bien manualmente los datos SNPs a analizar por la plataforma.

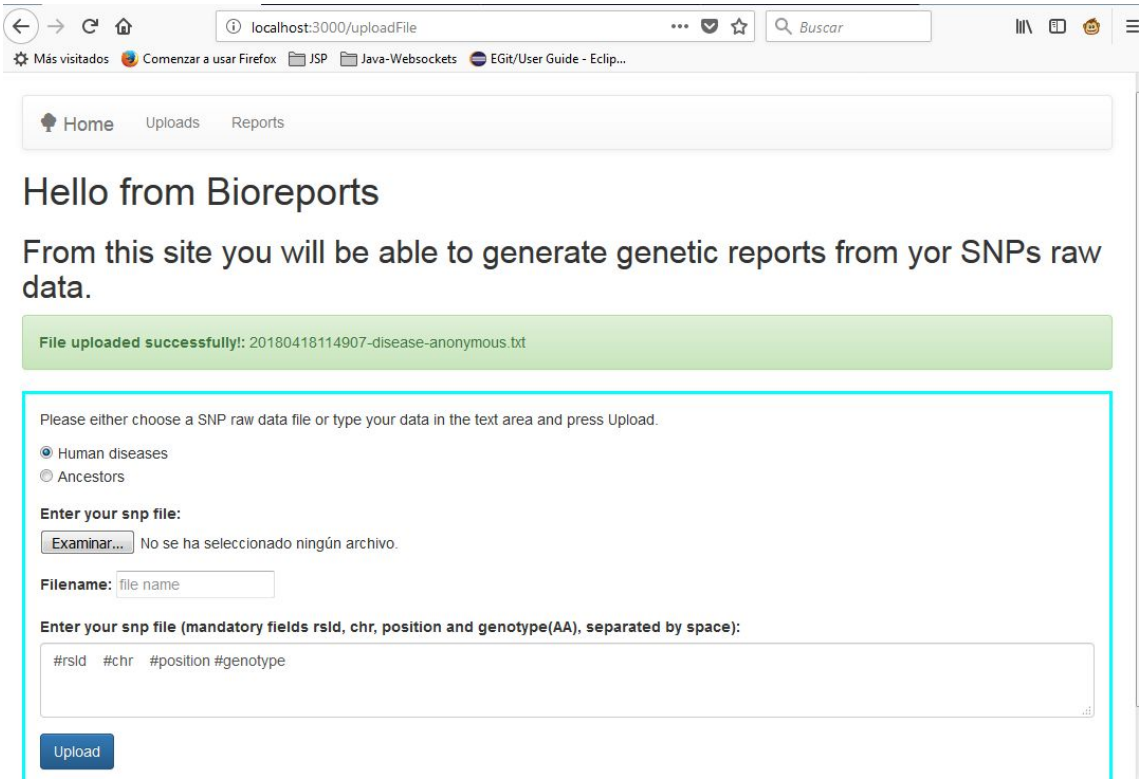
A continuación se muestran ilustraciones de esta pantalla:



Tal y como puede apreciarse, existen dos informes disponibles, “Human diseases” y “Ancestors”. Una vez seleccionado el tipo de informe se puede proporcionar la información de dos formas:

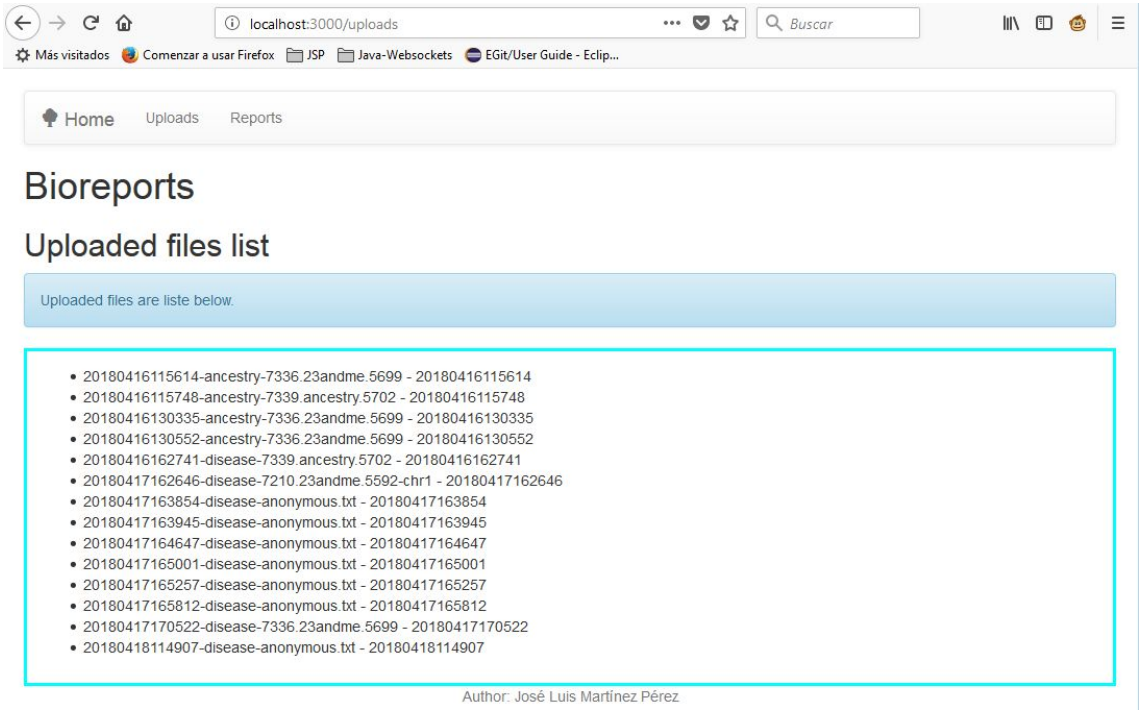
1. Subida de fichero de raw data.
2. Escritura manual de raw data con el formato que se explica en la caja de texto.

Una vez introducidos los datos del formulario, el proceso de análisis comienza al pulsar el botón “Upload”. Esta acción guarda el fichero original en el “Repositorio de raw data” y se mostrará información al usuario sobre si el fichero se ha subido correctamente o no a la plataforma.



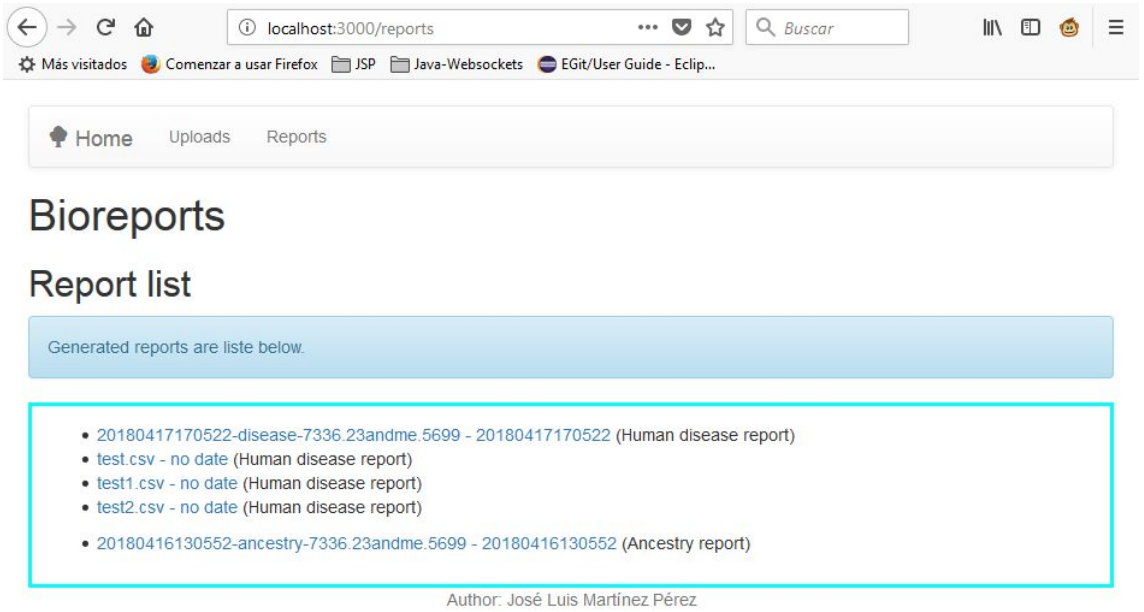
Listado de ficheros subidos.

Esta pantalla muestra la lista de ficheros que se han subido y que además han sido ya tratados para ajustar los datos al formato común de la plataforma y almacenados en el almacén de “Repositorio de datos transformados”.
Cada fichero viene identificado por la fecha en la que se ha generado, seguido del tipo de informe y el nombre original del fichero. Si los datos fueron introducidos manualmente, el nombre de fichero es “anonymous”.



Listado de informes.

La parte del listado de informes, muestra los informes ya finalizados del almacén de “informes generados” en forma de enlace. Al pinchar en cada enlace se accederá a la pantalla del detalle del informe.

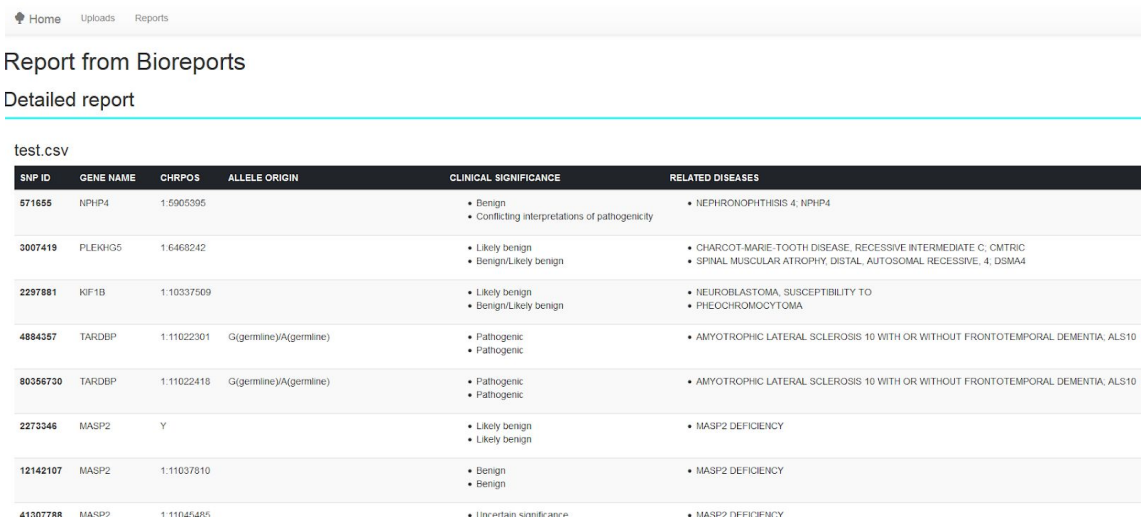


Detalle de informe.

Los informes disponibles en la primera versión de la plataforma son los informes de enfermedades humanas, incluyendo información nutrigenética y los de ancestros.

A continuación se muestran dos pantallazos con un ejemplo de cada uno de ellos.

Informe de enfermedades humanas:



SNP ID	GENE NAME	CHRPOS	ALLELE ORIGIN	CLINICAL SIGNIFICANCE	RELATED DISEASES
571655	NPHP4	1:5905395		• Benign • Conflicting interpretations of pathogenicity	• NEPHRONOPHTHISIS 4; NPHP4
3007419	PLEKHG5	1:6468242		• Likely benign • Benign/Likely benign	• CHARCOT-MARIE-TOOTH DISEASE, RECESSIVE INTERMEDIATE C; CMTRIC • SPINAL MUSCULAR ATROPHY, DISTAL, AUTOSOMAL RECESSIVE, 4; DSMA4
2297881	KIF1B	1:10337509		• Likely benign • Benign/Likely benign	• NEUROBLASTOMA, SUSCEPTIBILITY TO • PHEOCHROMOCYTOMA
4884357	TARDBP	1:11022301	G(germline)/A(germline)	• Pathogenic • Pathogenic	• AMYOTROPHIC LATERAL SCLEROSIS 10 WITH OR WITHOUT FRONTOTEMPORAL DEMENTIA; ALS10
80356730	TARDBP	1:11022418	G(germline)/A(germline)	• Pathogenic • Pathogenic	• AMYOTROPHIC LATERAL SCLEROSIS 10 WITH OR WITHOUT FRONTOTEMPORAL DEMENTIA; ALS10
2273346	MASP2	Y		• Likely benign • Likely benign	• MASP2 DEFICIENCY
12142107	MASP2	1:11037810		• Benign • Benign	• MASP2 DEFICIENCY
41307788	MASP2	1:11045485		• Uncertain significance	• MASP2 DEFICIENCY

Las columnas que se muestran son:

- SNP ID
- GENE NAME
- CHRPOS
- ALLELE ORIGIN
- CLINICAL SIGNIFICANCE
- RELATED DISEASES

Si el usuario tuviera algún SNP relacionado con algún gen implicado en los casos de nutrigenética estudiados en la plataforma, vería una ampliación de la pantalla anterior con la siguiente información:

En la imagen anterior se pueden ver algunos de los desórdenes o características que se han seleccionado como elementos del informe de nutrigenética y si se encuentra algún SNP relacionado con algún gen de un desorden, se marca y se especifica qué gen ha sido identificado.

En el ejemplo anterior se ha encontrado un SNP asociado a la deficiencia de sacarosa isomaltasa, SI.

Informe de ancestros:

Report from Bioreports

Detailed report

20180416130552-ancestry-7336.23andme.5699

REGION	REGION VALUE
Siberian	0.003640
Amerindian	0.000100
West_African	0.001940
Palaeo_African	0.000010
Southwest_Asian	0.049690
East_Asian	0.012750
Mediterranean	0.262020
Australasian	0.000250
Arctic	0.004010
West_Asian	0.108100
North_European	0.557470
South_Asian	0.000010
East_African	0.000010

Author: José Luis Martínez Pérez

En la imagen anterior se puede ver una columna con cada una de las 13 regiones contra las que se realiza el análisis y su valor.

2.4.2. Transformación de datos

Los ficheros de raw data tienen diferentes formatos, aunque sí que comparten algunos campos comunes. Además, al permitir la inserción de texto libre, hay que realizar alguna tarea de tratamiento de datos para conseguir homogeneizar en la medida de lo posible los datos que se tratarán en fases posteriores.
Además, este proceso inicial simplificará el tratamiento de datos en la generación de informes evitando comprobaciones de existencia y tipo de datos.

Se ha diseñado una tarea de ETL, la cuál:

- Elimina las lineas de comentarios (comienzan por #).
- Elimina los múltiples espacios en los separadores, así como las comas mediante expresioes regulares en lenguaje R:
    - Uno o más blancos ó comas (***[:blank:]|,+***)
    - Todos los espacios iniciales y finales (***^\\s+|\\s+$***)
- Guarda este resultado en el almacén de 'Datos Transformados'.

2.4.3. Consulta de datos y análisis de datos

El objetivo de este subaparatado es la descripción del proceso seguido para la obtención de los datos que se utilizarán en la generación de los informes.

Obtención de datos sobre enfermedades humanas

La búsqueda de datos se realizará en la base de datos NCBI, utilizando el paquete ***'rentrez'*** de R como herramienta de captura de datos.
Durante la realización de pruebas manuales con ficheros de test contra dbSNP de NCBI, se detectó que por una parte en los ficheros existen identificadores que no comienzan por "rs", sino por "i", y al buscar ese identificador en NCBI no aparecían resultados. En cambio, si ese mismo snp se buscaba por la tupla cromosoma:posición sí que se obtenían resultados.
De la misma forma, el caso opuesto también se ha detectado, es decir, la búsqueda por 'rsid' produce resultados pero al buscar por la tupla "cromosoma:valor" no se obtenían resultados y esto pasa porque la posición registrada en NCBI es diferente a la registrada en el fichero de raw data.
Este último caso de posiciones variantes se podría explicar por las razones siguientes:

- Debido al Linkage Disequilibrium (LD). A veces cuando variaciones en dos posiciones cercanas se heredan de forma conjunta se considera que es la misma variante, porque están en LD.
- Por otro lado, tenemos que tener en cuenta que el rs está basado en métodos de clustering sobre los alineamientos respecto al genoma de referencia. Así que si una posición se alinea con varias posiciones.Las razones oficiales por las que esto pueda ocurrir son :
    - La secuencia flanqueante del SNP es demasiado corta.
    - El SNP está en una región repetitiva del genoma.
    - Hay mucha variación en la secuencia flanqueante del SNP.

Por lo tanto, es posible que se pierdan algunos resultados en función de si se realiza una búsqueda u otra. Se ha parametrizado el procedimiento de búsqueda para poder realizar las búsquedas por ambos métodos (rsid y chr:position). En las pruebas realizadas se ha visto que **se obtienen muchos más resultados buscando por rsid**, por lo que es la opción por defecto.

- *search_string_by_chr_pos <- sprintf('%s[CHR] AND %s[CPOS] AND HUMAN[ORGN]', chr, pos)*
- *search_string_by_rsid <- sprintf('%s[RS] AND HUMAN[ORGN]', rsid)*

La solución correcta sería, que debido a la flexibilidad que nos ofrece la búsqueda en NCBI añadir complejidad a la cadena de búsqueda, pero también requeriría introducir complejidad en todos los procedimientos posteriores para tratar los datos devueltos.

- *search_string <- sprintf('%s[RS] OR (%s[CHR] AND %s[CPOS]) AND HUMAN[ORGN]', rsid, chr, pos)*

El flujo completo y el código fuente de consultas y procesado de datos se adjunta en el apartado de anexos pero a continuación se resaltan las partes más importantes:

- Construcción de la cadena inicial de búsqueda.
    - *search_string <- sprintf('%s[CHR] AND %s[CPOS] AND HUMAN[ORGN]', chr, pos)*
- Acceso a la base de datos SNP
    - *snp_search <- entrez_search(db="snp", term=search_term, retmax=20)*
- Extraer información de los SNPs
    - *snp_summary <- entrez_summary(db='snp',id = snp_search$ids)*
- De todos los resultados obtenidos, buscar si existen entradas en clinvar y extraer su significado clínico.
    - *clinvar_links <- entrez_link(dbfrom='snp', id=snp_search$ids, db='clinvar')*
    - *clinvar_summary <- entrez_summary(db='clinvar',id = clinvar_links$links$snp_clinvar)*
    - *clinvar_clinical <- extract_from_esummary(clinvar_summary, c("clinical_significance"))*
- Acceder a OMIM para obtener las enfermedades asociadas al SNP analizado.
    - *omim_links <- entrez_link(dbfrom = 'clinvar', id=clinvar_uids, db='omim')*
    - *omim_summary <- entrez_summary(db='omim',id = omim_links$links$clinvar_omim)*
    - *diseases <- extract_from_esummary(omim_summary, c("title"))*

No se han mostrado las partes del código para formar los data frames, filtrar datos innecesarios, etc.

Obtención de datos sobre poblaciones y ancestros

La estimación genética de ancestros en poblaciones humanas tiene aplicaciones importantes en estudios médicos. La genética de ancestros se usa como control en la estratificación de la población en estudios de asociación genética, y se utiliza para entender la base genética de la diferencia en susceptibilidad a enfermedades entre etnias y/o razas.

El estudio de ancestros puede dividirse en estimaciones locales y globales.
Las estimaciones locales se refieren a la identificación del origen ancestral de segmentos cromosómicos distintos dentro de un genoma individual, y estos métodos son un desarrollo más reciente en el campo. Las estimaciones globales buscan establecer proporciones ancestrales promediadas en todo el genoma de un individuo, de modo que las proporciones de cada ascendencia (que suman 1) puedan asignarse a cada individuo.

En el estudio previo de los algoritmos y software existente para realizar la clasificación de individuos en poblaciones se han identificado numerosas aproximaciones que tratan este problema.
En cuanto a los algoritmos que se utilizan, la literatura habla de dos aproximaciones principales.

- Clasificación no supervisada (STRUCTURE, ADMIXTURE).
- Análisis de componentes principales ó PCA en sus siglas en inglés (EIGENSTRAT).

Se ha seleccionado el algoritmo de ADMIXTURE, porque mejora el rendimiento respecto a STRUCTURE y por la documentación del software respecto a EIGENSTRAT.
Adicionalmente, se han visto más alternativas a la hora de analizar este informe, sobre todo paquetes de R. El problema de la mayoría de estos paquetes es el formato de entrada que necesitan y que requiere de subrutinas de transformación de datos en los formatos específicos de cada software.

Entre ellos destacan:

- SNPRelate
- CAnD
- AncestryMapper
- tutoRstructure
- GENESIS
- EthSEQ

Entre los formatos de entrada que necesitan estos paquetes, destacan:

- VCF
- PED y FAM (Generados mediante PLINK)
- BAM

Otro ejemplo es iAmix (https://bansal-lab.github.io/software/iadmix.html) una implementación en python que se decidió descartar por dos razones:

1. No estaba implementado con los lenguajes de programación seleccionados como base de la plataforma.
2. Se encontró un software compatible con los leguajes de programación de nuestra plataforma.

Finalmente, el paquete seleccionado para implementar este proceso es 'Radmixture'(https://cran.r-project.org/web/packages/radmixture/index.html).
Este paquete toma como entrada ficheros de texto que contienen el genotipo del individuo y lo compara con el modelo de referencia para poder realizar la clasificación. Este software se ajusta perfectamente al workflow que tenemos definido en nuestra plataforma, ya que tomará como entrada los ficheros subidos por parte del usuario, se realizarán los ajustes de formato necesarios y por último se analizará la estratificación de los SNPs en distintas poblaciones sin la necesidad de transformaciones adicionales o la utilización de otros programas o subrutinas, que introducen más complejidad a la solución.

En la bilbiografía, se exponene algunos workflows de análisis de ancestros utilizados por algunas empresas, como por ejemplo 23andme (https://blog.23andme.com/wp-content/uploads/2012/11/20121027_ancestry_painting_methods_poster.pdf)

En las líneas siguientes se muestran las partes claves de nuestro procedimiento de análisis de ancestros:

- Abrir fichero para analizar
  - *genotype <- read.table(file = path_to_file)*
- Cargar ficheros con datos de referencia con 13 posibles poblaciones
  - *load('/reference-data/globe13.alleles.RData')*
  - *load('/reference-data/globe13.13.F.RData')*
- Transformar el raw data al formato necesario para radmixture
  - *res <- tfrdpub(genotype, 13, globe13.alleles, globe13.13.F)*
- Calcular estadísticos
  - *ances <- fFixQN(resg, resq, res$f, tol = 1e-4, method = "BR", pubdata = "K13")*
- Mostrar resultados
  - *ances$q*

La imagen con los porcentajes de la composición de adn de este análisis se muestra a continuación:

```
> ances$q
       Siberian Amerindian West_African Palaeo_African Southwest_Asian East_Asian
result  0.00056      1e-05        1e-05          1e-05         0.12833     0.0036
       Mediterranean Australasian  Arctic West_Asian North_European South_Asian East_African
result       0.35092      0.00829 0.00785    0.20272        0.27918     0.01602       0.0025
> |
```

Filtrado de datos en Linkage Disequilibrium

“Linkage Disequilibrium (LD) es la asociación no aleatoria de alelos en diferentes loci en una población determinada. Se dice que los loci están en linkage disequilibrium cuando la frecuencia de asociación de sus diferentes alelos es mayor o menor de lo que se esperaría si los loci fueran independientes y se asociaran al azar.

El linkage disequilibrium está influenciado por muchos factores, incluida la selección, la tasa de recombinación, la tasa de mutación, la deriva genética, el sistema de apareamiento, la estructura de la población y la vinculación genética. Como resultado, el patrón de linkage disequilibrium en un genoma es una señal indicativa de los procesos genéticos de la población que lo están estructurando.

Además, puede existir un linkage disequilibrium entre alelos en diferentes loci sin ningún vínculo genético entre ellos e independientemente de si las frecuencias de los alelos están o no en equilibrio (no cambian con el tiempo).”
https://en.wikipedia.org/wiki/Linkage_disequilibrium

Como los ficheros de SNPs contienen muchos datos, para agilizar los cálculos se recomienda realizar una “poda o recorte” de datos que estén en LD, con ello obtendremos un conjunto de SNPs más “delgado”.
En la literatura, la herramienta de la que más citas se ha encontrado información para realizar un filtrado de SNPs en LD ha sido PLINK, pero como veremos más adelante presenta un problema que se ha resuelto realizando el cálculo de LD con el paquete SNPRelate de Bioconductor aprovechando los ficheros generados por PLINK.

El procedimiento para filtrar SNPs en LD con PLINK es el siguiente:

1. Transformar el fichero con los SNPs en formato .bed, .fam y .bim
    a. ***plink --23file test.txt --snps-only no-DI --make-bed --out bed_file***
2. Transformar el .bed en formato .ped (pedigrí)
    a. ***plink --bfile bed_file --make-founders --recode tab --out mypedfile***
3. Aplicar el filtro de LD con PLINK (https://www.cog-genomics.org/plink/1.9/ld)
    a. ***plink --file mypedfile --indep 50 5 2 --out pruned-snp-list***
    b. ***Ocurre un warning que impide la generación de los ficheros filtrados:***

.ped scan complete (for binary autoconversion).
Performing single-pass .bed write (609344 variants, 1 person).
--file: 23data/pruned-snp-list-temporary.bed +
23data/pruned-snp-list-temporary.bim + 23data/pruned-snp-list-temporary.fam written.
609344 variants loaded from .bim file.
1 person (0 males, 1 female) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1 founder and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.977875.
609344 variants and 1 person pass filters and QC.
Note: No phenotypes present.
Warning: Skipping --indep since there are less than two founders.
(--make-founders may come in handy here.)

4. Utilizamos el filtro sugerido --make-founders (https://www.cog-genomics.org/plink/1.9/filter#make_founders), pero seguimos teniendo el mismo error.
    a. ***plink --23file test.txt --make-founders require-2-missing first --snps-only no-DI --make-bed --out bed_file***

Utilizando PLINK se pueden realizar transformaciones entre distintos formatos, por ejemplo, tomando como entrada un fichero con formato 23andme podemos generar uno en formato VCF fácilmente:

plink --23file 23data/7210.23andme.5592.txt --snps-only no-DI --recode vcf --out 23data/out.vcf

// Fichero 7210.23andme.5592.txt

```
# rsid  chromosome  position  genotype
rs548049170  1  69869  TT
rs13328684  1  74792  --
rs9283150  1  565508  AA
i713426  1  726912  --
rs116587930  1  727841  GG
```

// Fichero out.cvf

```
##fileformat=VCFv4.2
##fileDate=20180422
##source=PLINKv1.90
##contig=<ID=1,length=249222528>
##contig=<ID=2,length=243178151>
##contig=<ID=3,length=197884213>
##contig=<ID=4,length=191027349>
##contig=<ID=5,length=180715141>
##contig=<ID=6,length=170904809>
##contig=<ID=7,length=159124174>
##contig=<ID=8,length=146300856>
##contig=<ID=9,length=141101940>
##contig=<ID=10,length=135473015>
##contig=<ID=11,length=134945121>
##contig=<ID=12,length=133838354>
##contig=<ID=13,length=115106997>
##contig=<ID=14,length=107283151>
##contig=<ID=15,length=102462480>
##contig=<ID=16,length=90239298>
##contig=<ID=17,length=81151540>
##contig=<ID=18,length=78010621>
##contig=<ID=19,length=59097934>
##contig=<ID=20,length=62954926>
##contig=<ID=21,length=48099611>
##contig=<ID=22,length=51214797>
##contig=<ID=23,length=155234708>
##contig=<ID=24,length=58997093>
##contig=<ID=26,length=16527>
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genome">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  FAM001_ID001
1  69869  rs548049170  T  .  .  .  PR  GT  0/0
1  74792  rs13328684  N  .  .  .  PR  GT  ./.
1  565508  rs9283150  A  .  .  .  PR  GT  0/0
1  726912  i713426  N  .  .  .  PR  GT  ./.
```

Tal y como se comentaba en párrrafos anteriores, aprovechado la herramienta PLINK, podemos generar los formatos .bed, .fam y .bim que se utilizan como entrada del paquete SNPRelate de Bioconductor (http://corearray.sourceforge.net/tutorials/SNPRelate/ ver apartado “LD-based

SNP pruning"), para luego poder aplicar su función de LD y obtener un listado de SNPs sin LD. Posteriormente, nos quedaremos con el subconjunto de esos SNPs en nuestro data frame original para acelerar los cálculos.

Los pasos detallados de cómo funcciona el algooritmo de poda de SNPs en LD se pueden encontrar en la doumentación del paquete de Bioconductor (https://www.rdocumentation.org/packages/SNPRelate/versions/1.6.4/topics/snpgdsLDpruning). A continuación se exponen estos pasos en lenguaje pseudo-algorítmico obtenidos del enlace anterior en inglés:

1. *Randomly select a starting position i, and let the current SNP set S = { i };*
2. *For each right position j from i+1 to n: if any LD between j and k is greater than ld.threshold, where k belongs to S, and both of j and k are in the sliding window, then skip j; otherwise, let S be S + { j };*
3. *For each left position j from i-1 to 1: if any LD between j and k is greater than ld.threshold, where k belongs to S, and both of j and k are in the sliding window, then skip j; otherwise, let S be S + { j };*
4. *Output S, the final selection of SNPs.*

Es muy importante mencionar los problemas que se han identificado y las decisiones que se han tomado sobre la aplicación este filtro de LD.
El principal problema es que según qué paquete de software se utilizaba para su cálculo se producían dos efectos que limitaban mucho el informe de enfermedades humanas.
Por una parte, utilizando "SNPRelate" se filtraban muchos SNPs que sin ese filtro sí que arrojaban resultados positivos sobre posible significado clínico, además reducían el tamaño del juego de datos SNP de un promedio de 650k SNPs a 5k SNPs.
Por otra parte, utilizando "ProxySNP" a modo de prueba de concepto para contrastar resultados contra "SNPRelate", el efecto que se producía era que al realizar llamadas por red, el proceso de filtrado aplicado desde un PC portátil estándar se dilataba muchísimo en el tiempo, del orden de 48 horas en filtrar los 650k SNPs, aún filtrando menos resultados que "SNPRelate" y por lo tanto siendo más fiable, esta solución no es viable con las condiciones de cálculo y red en el PC disponible para el desarrollo del proyecto, siendo un candidato a formar parte de la solución para un entorno de producción con capacidades de cálculo y red mucho mayores.
Con estas pruebas y los resultados arrojados, para el informe de enfermedades se decidió no aplicar el filtro y así poder tener la máxima información sobre enfermedades humanas.

Resumen de experimentos realizados y justificación de decisiones tomadas

Por capacidad de computación en el pc portátil de pruebas, se han dividido los ficheros por cromosomas. Se han hecho pruebas con ficheros públicos u 'open data' de distintas empresas, tales como *23andme* y *ancestry*, así como chips *illumina* también 'open data'. En las siguientes tablas se muestran una comparativa sobre los resultados que se pierden al aplicar el software seleccionado sobre varios ficheros de datos SNP.

a. Análisis del fichero 7339.ancestry.5702

El análisis se ha divido en 2 tablas porque para poder procesar el fichero sin filtro se ha tenido que hacer en varias ejecuciones porque por ejemplo, el cromosoma 1 ha tardado 8 horas aproximadamente en procesarse completamente con el PC en el que se realizan los experimentos.

Fichero	SNPs original	Tiempo de proceso con filtro LD	Enfermedades sin filtro LD (Pathogenic / Detectadas)
7339.ancestry.5702-chr1	49.513	~ 8 horas	203 / 395
7339.ancestry.5702-chr2	51.094	~ 8 horas	0 / 1
7339.ancestry.5702-chr3	40.130	~ 8 horas	67 / 170
7339.ancestry.5702-chr4	35.289	~ 8 horas	56 / 120
7339.ancestry.5702-chr5	37.654	~ 8 horas	39 / 235
7339.ancestry.5702-chr6	40.135	~ 8 horas	71 / 212
7339.ancestry.5702-chr7	34.066	~ 8 horas	13 / 30

Fichero	SNPs original	SNPs después del filtro LD	Tiempo de proceso con filtro LD	Enfermedades con filtro LD (Pathogenic / Detectadas)
7339.ancestry.5702	650.400	4.803	~ 2 horas	1 / 15

De esta forma podemos ver que el cromosoma 1 de este fichero consta de 49.513 SNPs, de los cuales 395 tienen un significado clínico y de éstos, 203 se relacionan con enfermedades. La misma explicación para las demás filas de la tabla.
Al aplicar la poda o filtro de LD, los SNPs que quedan en el cromosoma 1 son 4.803, estando relacionados con enfermedades 1 de 15 con significado clínico.

b. Análisis del fichero 7210.23andme.5592

De la misma forma que el fichero anterior, el análisis se ha dividido en 2 tablas, una sin filtro LD y otra con filtro LD.

Fichero	SNPs original	Tiempo de	Enfermedades

		proceso sin filtro LD	sin filtro LD (Pathogenic / Detectadas)
7210.23andme.5592-chr1	48.312	~ 8 horas	29 / 79
7210.23andme.5592-chr2	51.771	~ 8 horas	42 / 230
7210.23andme.5592-chr3	43.023	~ 8 horas	21 / 89
7210.23andme.5592-chr4	37.289	~ 8 horas	56 / 119

Fichero	SNPs original	SNPs después del filtro LD	Tiempo de proceso con filtro LD	Enfermedades con filtro LD (Pathogenic / Detectadas)
7210.23andme.5592	638.469	4.762	~ 2 horas	0 / 12

Al igual que en la tabla anterior, el cromosoma 1 de este fichero consta de 48.312 SNPs, de los cuales 79 tienen significado clínico y de éstos, 29 se relacionan con enfermedades. Lo mismo sucede en las demás filas de la tabla. Al aplicar la poda o filtro de LD, los SNPs que quedan en el cromosoma 1 son 4.762, estando relacionados con enfermedades 12 y entre éstos, no se detectan SNPs patogénicos.

c. Conclusión

Viendo los resultados de las dos tablas anteriores se concluye que aplicando el filtro se pierden muchísimos SNPs patogénicos que no se van a poder mostrar en el informe y es por esto que se descarta aplicar este filtro para el informe de enfermedades.

2.4.4. Generación de informes

Esta tarea, forma parte del módulo de front-end y consiste en una serie de subrutinas para ofrecer la visualización de los informes finales generados en apartados anteriores desde la aplicación de la plataforma web. Esta tarea, básicamente realiza procesado y formateado de datos, para por ejemplo mostrar en forma de lista HTML los nombres de las enfermedades asociadas a cada SNP, en caso de que exista más de una.

Básicamente, el proceso obtiene el informe depositado en el almacén de informes generados, con formato de fichero csv (delimitado por comas ‘,’) y en

función del tipo de informe se trata la información del mismo y se muestra por pantalla.

2.5 Nutrigenética

Este informe está relacionado con el de enfermedades, el procedimiento es similar ya que está basado en la obtención de información sobre genes involucrados en la interacción del organismo con la ingesta de alimentos.
Por ejemplo, genes que están relacionados la tolerancia a la fructosa y lactosa, dificultad a la hora de pérdida de peso, metabolismo de la cafeína y su implicación en riesgo de ataque cardíaco, metabolismo de los ácidos grasos, de los carbohidratos, incluso en la prevención de lesiones deportivas.

Por lo tanto, se van a preseleccionar algunos de los genes identificados en el desarrollo de procesos metabólicos alimentarios y para cada SNP se comprobará que no está relacionado tales genes.

La siguiente tabla muestra los procesos metabólicos implicados en aspectos de la nutrición y la lista de genes asociados a cada uno de ellos.

Desorden alimentario	**Genes implicados**
Fatty acid metabolism	• APOA2 • LEPR • PPARA • ADIPOQ • ADIPOA5_1 • ADIPOA5_2
Carbohydrates metabolism	• MTNR1B • G6PC2 • FADS1 • GCK • ADRA2A • CRY2 • GCKR • PROX1 • MADD • DGK • TMEM • TCF7L2 • SLC30A8 • CRY2 • GLIS3 • ADCY5

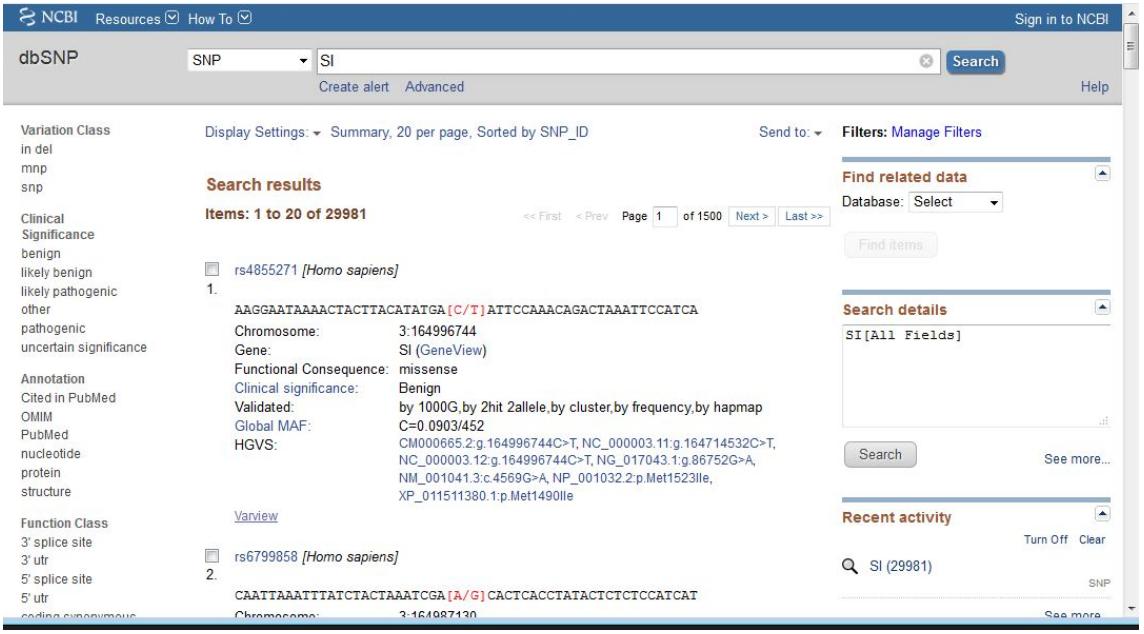
Salt sensivity	• AGT • ACE
Nutrigenetics & Omega 3/6	• APAO5 • FADS1
Folate metabolism	• MTR • MTRR • MTHFR • MTHFR_1 • MTHFR_2
Injury Prevention	• COL1A1
Lactose Intolerance risk	• MCM6
Testosterone optimisation	• CYP19 • HSD11B1 • ACTN3 • HSD11B1
Anti-inflammatory nutrients	• CRP • IL6 • GSTP1 • TNF_Alpha
Vitamin profile	• FUT2NBPF3 • FUT2 • CYP26B1 • GC • CYP2R1_1 • CYP2R1_2 • NBPF3 • BCM01_1 • BCM01_2 • SLC23A1 • SLC23A2
Fat loss response to green tea	• COMT
Fat loss	• FTO • ADRB2 • PPARG • PGC1-alpha • TFAM
Response to chromium picolinate	• DRD2
Caffeine metabolism	• ADRA2A • CYP1A2

Poor sleep on increased consumption of caffeine.	● ADORA2A
Increased risk of fatty liver on low choline intake.	● MTHFD1
glucose-6-phosphate dehydrogenase deficiency, a condition which is mostly triggered by the consumption of fava beans.	● G6PD
Predisposed to Hemochromatosis when their iron intake is high.	● HFE
Sucrase-isomaltase deficiency when their sucrose intake is high.	● SI
Hereditary fructose intolerance when their fructose intake is high.	● ALDOB
Enhanced alcohol dehydrogenase activity and enables them to have more drinks as compared to non-carriers.	● ALDH
Low copy numbers of the gene and reduced ability to digest starch.	● AMY1
Increased bitter taste perception and decreased consumption of vegetables.	● TAS2R38

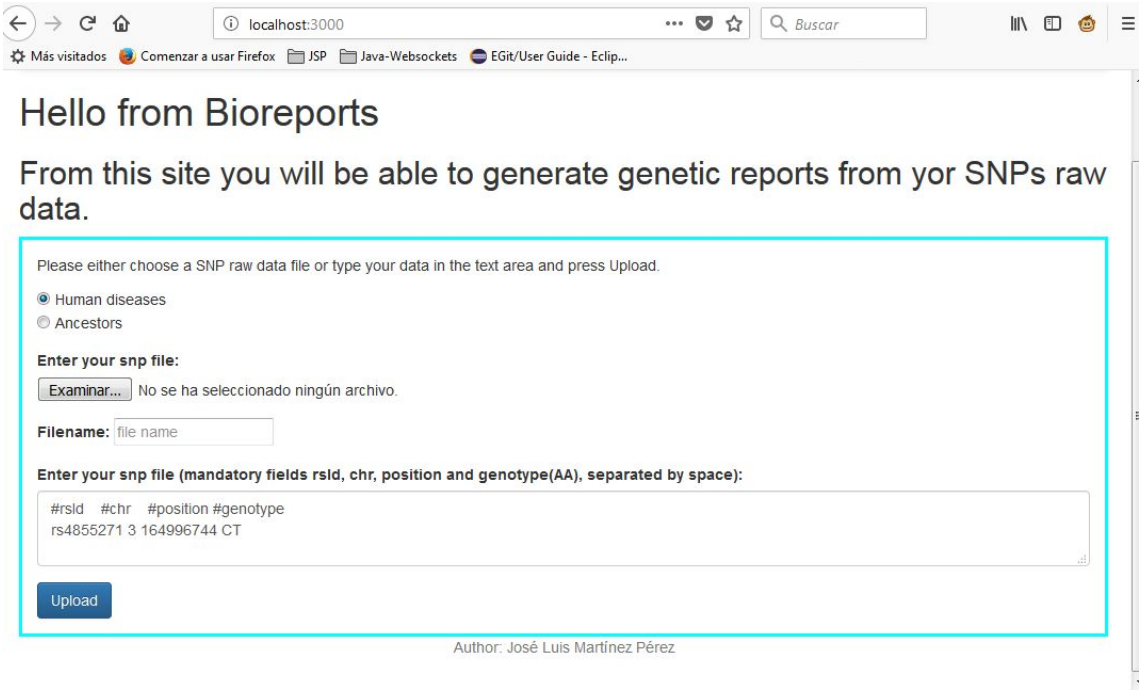
Sobre este informe, hay que hacer notar que los genes listados son un subconjunto muy reducido del total de genes que tenemos y por lo tanto es difícil encontrar un juego de datos que una vez procesado tenga alguno de los SNPs que se corresponden con los genes anteriores.

Por este motivo, para realizar el testing de la funcionalidad desarrollada, se han añadido “manualmente” SNPs relacionados con estos genes obtenidos desde la base de datos dbSNP del NCBI a ficheros ya existentes. También es posible introducir SNPs conocidos desde el formulario de entrada de datos de texto, en lugar de subir un fichero para su análisis.

Por ejemplo, para localizar un SNP relativo a “Sucrase-isomaltase deficiency” (SI), buscamos en NCBI y podemos encontrar varios SNPs:



Con esta información sobre el rsId, cromosoma y posición, podríamos introducir manualmente este SNP desde la aplicación y veríamos el resultado en un informe sobre un único SNP a modo de prueba.



Report from Bioreports

Detailed report

20180417170522-disease-7336.23andme.5699

Diseases

SNP ID	GENE NAME	CHRPOS	ALLELE ORIGIN	CLINICAL SIGNIFICANCE	RELATED DISEASES
4855271	SI	1:212895253	T(germline)/C(germline)	• Likely benign • Benign/Likely benign	• Sucrase-isomaltase deficiency; SI

Nutrigeomics

NUTRITION STUDY	GENES DETECTED
fatty_acid_metabolism	None
carbohydrates_metabolism	None
fatty_liver	None
glucose-6-phosphate_dehydrogenase_deficiency	None
hemochromatosis	None
sucrase-isomaltase_deficiency	☑ SI
hereditary_fructose_intolerance	None
high_alcohol_tolerance	None

3. Conclusiones

3.1. Conclusiones del trabajo

La conclusión principal es que de forma general el trabajo ha sido muy satisfactorio, llegando a cubrir la totalidad de los objetivos marcados en el inicio del proyecto.

El proyecto ha tenido varias dificultades de diferente ámbito que se han ido solventando mediante soluciones equilibradas tanto en la calidad de la solución como en el tiempo de desarrollo de la misma.

Las dificultades más destacables se agrupan en dificultades técnicas, dificultades conceptuales, teóricas y de interpretación de datos, dificultades de selección de software, tratamiento y formato de datos.

En cuanto a dificultades técnicas, destaca el hecho de que este tipo de procesos requiere de una capacidad de procesamiento de datos bastante elevada, por lo que trabajar en computadoras domésticas ralentiza bastante el tratamiento y generación de datos. También debe hacerse notar que cada proceso o informe que se quiera implementar necesita un estudio teórico y técnico sobre cómo va a realizarse. Un informe puede requerir muchos accesos a bases de datos para obtener información, mientras que otro puede necesitar mucho cálculo de datos.

En lo relativo a conceptos, teoría e interpretación de datos, lo más destacable es la dificultad de interpretar correctamente la numerosa información que existe en las bases de datos, cómo extraer la información correcta y necesaria para cada tipo de informe y/o consulta que se requiera. También requirió de bastante tiempo de estudio la forma de acceder a las bases de datos, ya que son conjuntos de bases de datos enlazadas entre sí, y hay que ver qué información se puede extraer de una para poder consultar en otra, conformando que la información que se obtiene es correcta. La primera cuestión que se tuvo que resolver sobre NCBI fue, ¿por dónde se empieza a consultar esta base de datos tan enorme?. Si buscamos enfermedades humanas por ejemplo, ¿es mejor comenzar por OMIM para listar enfermedades y sacar todos los SNPs asociados a cada una de ellas y luego acceder a clinvar y dbSNP para ampliar la información? ¿ó es mejor comenzar por dbSNP para cada uno de nuestros SNP de raw data e ir obteniendo información de clinvar, OMIM, etc?. Para contestar a estas cuestiones la única solución es estudiar la documentación disponible de NCBI, y realizar pequeñas pruebas de concepto en las que accediendo a las bbdd de NCBI se pueden extraer conclusiones. En este caso, la búsqueda de información se realizó pinalmente desde dbSNP, pasando por Clinvar y finalmente por OMIM para ver la enfermedad con la que estaba asociada cada snp.

A la hora de seleccionar el software, también existe gran diversidad de opciones, bases de datos, herramientas tanto de escritorio como online que dificultan la tarea de elegir las herramientas adecuadas para el desarrollo del proyecto. En este punto es importante acudir a la literatura existente para por lo menos tener una idea de lo que se puede llegar a utilizar en proyectos similares. La conclusión básica de este punto es que hay que tener claro qué se quiere y luego hay que explorar cada herramienta para ver sus posibilidades y ver qué necesita cada herramienta como entrada y qué produce como salida.

Por ejemplo, hasta encontrar los paquetes y herramientas que se han utilizado, las herramientas alternativas como STRUCTURE, ADMIXTURE, PLINK, paquetes de Bioconductor, etc requería cada una de ellas un tipo diferente de formato (.ped, .bed, .fam, .bim. bam, .vcf). A su vez, para generar estos ficheros partiendo de nuestro raw data de openSNP, era necesario realizar más transformaciones de datos, lo que incrementaba tanto la dificultad técnica como el sobrecoste computacional.

3.2. Consecución de objetivos

Recordando el objetivo principal del proyecto, el diseño e implementación de una plataforma de generación de informes médicos personalizados en la que partiendo de datos SNPs se ofrezca información genética al usuario, éste, se puede considerar como cumplido ya que se ha desarrollado una plataforma web con la posibilidad de ofrecer dos tipos diferentes de informes, por un lado los médicos y de enfermedades y por otro, de composición de ancestros.

De los objetivos secundarios que estaban de alguna forma ligados al objetivo principal, se han cumplido los objetivos teóricos de profundización en conceptos genéticos y pipelines de análisis genéticos existentes, y técnicos, como son el desarrollo con R y el uso de librerías bioinformáticas y bioestadísticas, el acceso automatizado a bases de datos y el tratamiento y transformación de datos.

El único objetivo marcado inicialmente aunque de forma opcional y que no se ha desarrollado de forma práctica ha sido el O5, estudio de interacción entre genes para producir enfermedades humanas debido a su complejidad.

3.3. Cumplimiento de planificación

La planificación se ha cumplido en cada uno de los entregables de los que constaba el trabajo de fin de máster. Sí que cabe mencionar que ha habido tareas con desvíos en cuanto a tiempos en relación a la planificación realizada al inicio. Estos desvíos han sido positivos (llevado menos tiempo que el planificado) en algunas tareas y en otras han sido negativos (más tiempo que el estimado), por lo que de forma general, se han compensado y se ha podido cumplir la planificación establecida.

3.4. Líneas de trabajo futuro

Como líneas de trabajo futuro, y con el objetivo de convertir el prototipo de la plataforma en un producto que pueda servir como herramienta comercial, se proponen las siguientes tareas:

- Mejorar la gestión de errores, informando al administrador de la plataforma sobre si ha habido errores durante la generación de informes en algún SNP concreto, etc.
- Mejorar el tratamiento heterogéneo de datos desde NCBI, ya que algunos registros no vienen con el formato esperado y aún aportando información se descartan al generar un error en el proceso de interpretación de datos.
- Ampliar la oferta de informes disponibles.
- Integrar más bases de datos con información ómica, como openSNP.
- Más detalle en los informes, aportando más información que la que se ofrece actualmente, por ejemplo, mostrando una gráfica de la cadena de ADN con los snps tal y como se hace desde la web de dbSNP.

4. Glosario

Acrónimo	Significado
NCBI	National Center for Biotechnology Information
SNP	Single nucleotide polymorphism
OP	Objetivo Principal
O1	Objetivo secundario 1
O2	Objetivo secundario 2
O3	Objetivo secundario 3
O4	Objetivo secundario 4
O5	Objetivo secundario 5
LD	Linkage Disequilibrium

5. Bibliografía

Empresas informes genéticos

Empresa	URL	Fecha de visita
23andme	https://www.23andme.com	07/03/2018
Tellmegen	https://www.tellmegen.com	07/03/2018
24genetics	https://24genetics.com	12/03/2018
FamilyTreeDNA	https://www.familytreedna.com	12/03/2018
AncestryDNA	https://www.ancestry.com	12/03/2018
deCODE	https://www.decode.com	19/03/2017

Medicina personalizada

URL	Fecha de visita
https://www.medicinapersonalizadagenomica.com	07/03/2018
http://www.genealogia.org.mx/molecular/index.php?option=com_content&task=view&id=31&Itemid=2	12/03/2018
https://www.elconfidencial.com/tecnologia/ciencia/2017-03-19/genealogia-adn-genetica-suecia-murcia-ancestros-vikingos_1350698/	12/03/2018
https://en.wikipedia.org/wiki/Genealogical_DNA_test	12/03/2018
http://www.roche.es/Sobre_Roche/medicina-personalizada.html	12/03/2018

Genética

URL	Fecha de visita

https://ghr.nlm.nih.gov/primer/genomicresearch/snp	13/03/2018
https://es.wikipedia.org/wiki/Gen%C3%A9tica_humana	13/03/2018
https://www.etilmercurio.com/em/que-tan-grande-es-tu-genoma-cabe-en-un-pendrive/	14/03/2018
https://www.xatakaciencia.com/genetica/las-cifras-mas-curiosas-del-adn	14/03/2018
https://en.wikipedia.org/wiki/Genome-wide_association_study	19/03/2017
https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37	19/03/2017

Datos SNPs

URL	Fecha de visita
https://opensnp.org	19/03/2018
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0089204	19/03/2018
http://samtools.github.io/hts-specs/VCFv4.3.pdf	19/03/2018
https://academic.oup.com/bioinformatics/article/27/15/2156/402296	19/03/2018
https://en.wikipedia.org/wiki/Variant_Call_Format	19/03/2018
https://faq.myheritage.com/DNA/MyHeritage-DNA-Test/951697341/How-should-I-interpret-my-raw-DNA-data.htm	10/04/2018

Temática	URL	Fecha de visita
Ancestry Informative Marker (AIM)	• https://en.wikipedia.org/wiki/Ancestry-informative_marker	10/04/2018

	• https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3433799/ • https://www.ucl.ac.uk/mace-lab/debunking/understanding	
ANCESTRY WORKFLOW	• https://blog.23andme.com/wp-content/uploads/2012/11/20121027_ancestry_painting_methods_poster.pdf • https://botany.natur.cuni.cz/hodnocenidat/Lesson_05_tutorial.pdf • https://onlinelibrary.wiley.com/doi/full/10.1002/sim.6605 • http://blogs.discovermagazine.com/gnxp/2011/03/analyzing-ancestry-with-admixture-step-by-step/	14/04/2018
Articles and books	• https://www.palgrave.com/gp/book/9783319628806 • https://www.ncbi.nlm.nih.gov/pubmed/28334108 • https://academic.oup.com/bioinformatics/article-abstract/33/14/2148/3002763?redirectedFrom=fulltext • https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5633392/ • https://www.nature.com/articles/ncomms4513 • https://www.sciencedirect.com/science/article/pii/S0002929712004661 • https://www.sciencedirect.com/science/article/pii/S1872497316300667 • https://www.nature.com/articles/ncomms4513	14/04/2018
Software	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3542037/	10/04/2018
STRUCTURE	https://www.sciencedirect.com/science/article/pii/S1872497314000039	14/04/2018
EIGENSTRAT	• https://github.com/DReichLab/EIG	14/04/2018

	• https://github.com/DReichLab/EIG/blob/master/EIGENSTRAT/README • https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3215268/	
ADMIXTURE	• https://www.genetics.ucla.edu/software/admixture/ • https://discovery.illumina.com/variant-discovery/admixture-mapping-and-common-variants-4?sciid=2017318ILK1 • http://gaworkshop.readthedocs.io/en/latest/contents/07_admixture/admixture.html • https://www.biostars.org/p/255121/ • https://www.biostars.org/p/231885/ • http://blogs.discovermagazine.com/gnxp/2011/03/analyzing-ancestry-with-admixture-step-by-step/	14/04/2018
Mejoras Admixture	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-246	14/04/2018
PLINK	• https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/ • https://www.staff.ncl.ac.uk/heather.cordell/pak2010MDS.html • https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&cad=rja&uact=8&ved=2ahUKEwiJp7nK07baAhVGpZQKHfchA6AQFjADegQIABBP&url=http%3A%2F%2Fwww.ru.nl%2Fpublish%2Fpages%2F669642%2Fcongenomics_plink_tutorial_davey.pdf&usg=AOvVaw25lEKjOaso2bYze0_-zGG5	

	• http://blogs.discovermagazine.com/gnxp/2013/01/using-your-23andme-data-in-plink/ • http://zzz.bwh.harvard.edu/plink/tutorial.shtml • https://anthrogenica.com/showthread.php?9812-PLINK-and-23andme-raw-data-file • http://www.jade-cheng.com/au/23andme-to-plink/ • https://leesjohn.wordpress.com/2014/03/18/impute-your-whole-genome-from-23andme-data/	
R packages	• https://cran.r-project.org/web/packages/AncestryMapper/vignettes/AncestryMapper2.0.html • http://membres-timc.imag.fr/Olivier.Francois/tutoRstructure.pdf • https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-317 • https://bioconductor.org/packages/3.7/bioc/vignettes/GENESIS/inst/doc/pcair.html • https://bioconductor.org/packages/release/bioc/vignettes/SNPRelate/inst/doc/SNPRelateTutorial.html • https://cran.r-project.org/web/packages/EthSEQ/vignettes/EthSEQ.html	14/04/2018
Radmixture package	• https://cran.r-project.org/web/packages/radmixture/index.html • https://github.com/wegene-llc/radmixture/blob/master/README.md • https://cran.r-project.org/web/packages/radmixture/radmixture.pdf	14/04/2018

iAdmix software	• https://bansal-lab.github.io/software/iadmix.html • https://github.com/vibansal/ancestry	14/04/2018
Transforming data	• https://samtools.github.io/bcftools/howtos/convert.html • https://samtools.github.io/bcftools/howtos/install.html • https://www.biostars.org/p/255121/ • https://www.biostars.org/p/132940/ • http://augustogarcia.me/statgen-esalq/Hapmap-and-VCF-formats-and-its-integration-with-onemap/ • https://www.biostars.org/p/152300/ • https://indo-european.eu/human-ancestry/merge-remove-convert-datasets-bed-ped-fam-geno-snp-plink/ • https://www.researchgate.net/post/How_do_I_convert_a_SNP_genotype_table_into_plink_binary_PED_files	14/04/2018
Analysing DNA sequencing data	• https://davetang.org/muse/2015/07/24/dna-sequencing-data/	05/05/2018

Linkage Desequilibrium

URL	Fecha de visita
https://en.wikipedia.org/wiki/Linkage_disequilibrium	07/03/2018
https://www.sciencedirect.com/topics/neuroscience/linkage-disequilibrium	07/03/2018
https://www.ndsu.edu/pubweb/~mcclean/plsc731/Linkage%20Disequilibrium%20-%20Association%20Mapping%20in%20Plants-lecture-overheads.pdf	07/03/2018

https://www.rdocumentation.org/packages/SNPRelate/versions/1.6.4/topics/snpgdsLDpruning	29/04/2018

Nutrigenómica y nutrigenética

URL	Fecha de visita
https://www.anabolicgenes.com/en/nutrition.html	29/04/2018
https://www.xcode.in/dna-and-nutrition/nutrigenomics-dna-testing-dna-diet-nutrigenetics	29/04/2018
https://24genetics.com/en/nutrigenetics	29/04/2018
https://www.theguardian.com/lifeandstyle/2016/feb/29/die-now-diet-later-could-nutrigenetics-save-your-life	29/04/2018
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5330198/	29/04/2018
http://www.saragottfriedmd.com/%E2%80%8Efive-genes-that-make-it-hard-to-lose-weight-and-what-you-can-do-to-combat-them/	29/04/2018

Bases de datos

Base de datos	URL	Fecha de visita
Entrez	https://www.ncbi.nlm.nih.gov/gquery/	19/03/2018
dbSNP	https://www.ncbi.nlm.nih.gov/snp/	19/03/2018
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/	19/03/2018
OMIM	https://www.ncbi.nlm.nih.gov/omim/	19/03/2018

dbSNP book	https://www.ncbi.nlm.nih.gov/books/NBK174586/	19/03/2018
SNPedia	https://www.snpedia.com/index.php/SNPedia	07/04/2018
Ensembl	http://www.ensembl.org	07/04/2018
SNPsnap	https://data.broadinstitute.org/mpg/snpsnap/index.html	07/04/2018

Software

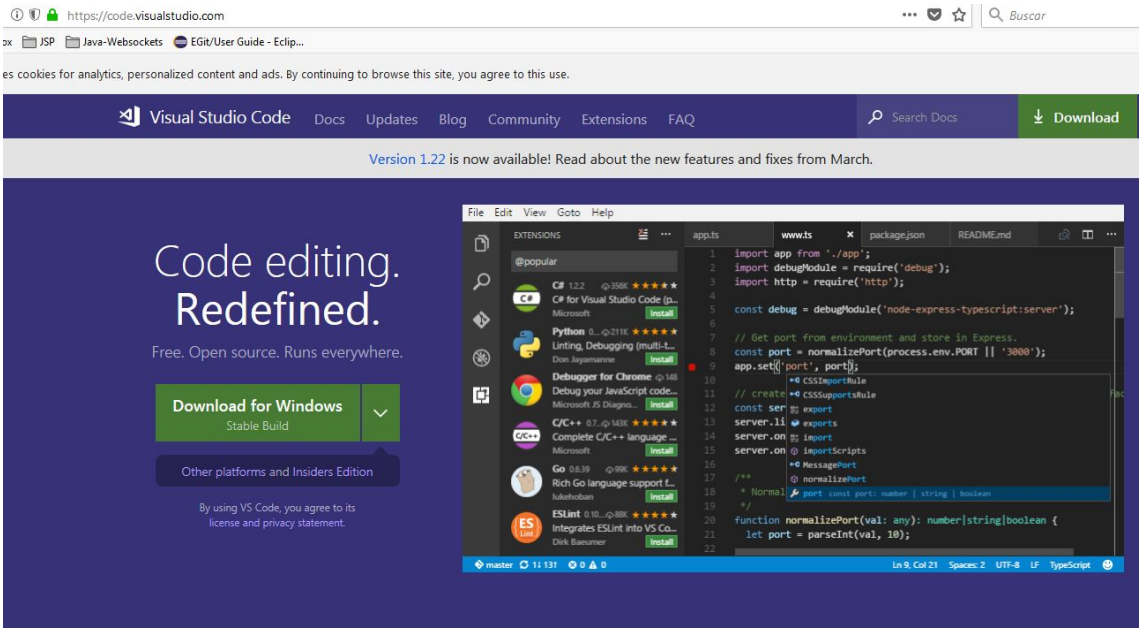
Software	URL	Fecha de visita
Bioconductor	https://www.bioconductor.org	07/03/2018
reutils	• https://github.com/gschofl/reutils • https://cran.r-project.org/web/packages/reutils/reutils.pdf • https://www.rdocumentation.org/packages/reutils/versions/0.2.2/topics/reutils	03/04/2018
rentrez	https://cran.r-project.org/web/packages/rentrez/vignettes/rentrez_tutorial.html	03/04/2018

6. Anexos

6.1 Anexo I. Instalación del entorno de desarrollo.

Instalación de Visual Studio Code (VS Code).

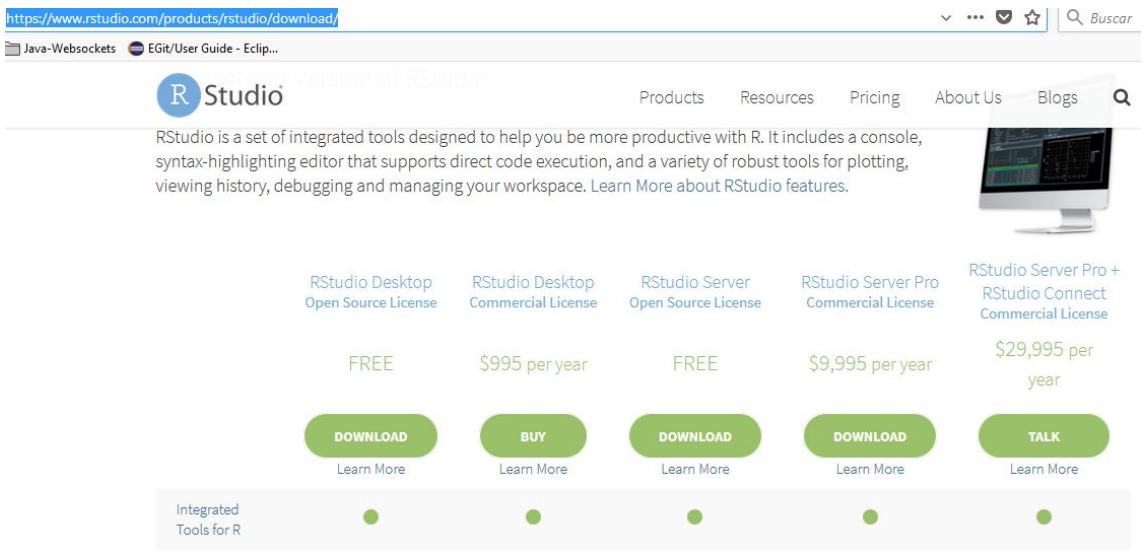
Visual Studio Code se puede descargar de la siguiente URL (https://code.visualstudio.com/)



Una vez descargado, sólo hay que ejecutar el instalador y seguir el asistente para completar la instalación.

Instalación de RStudio, R y RScript.

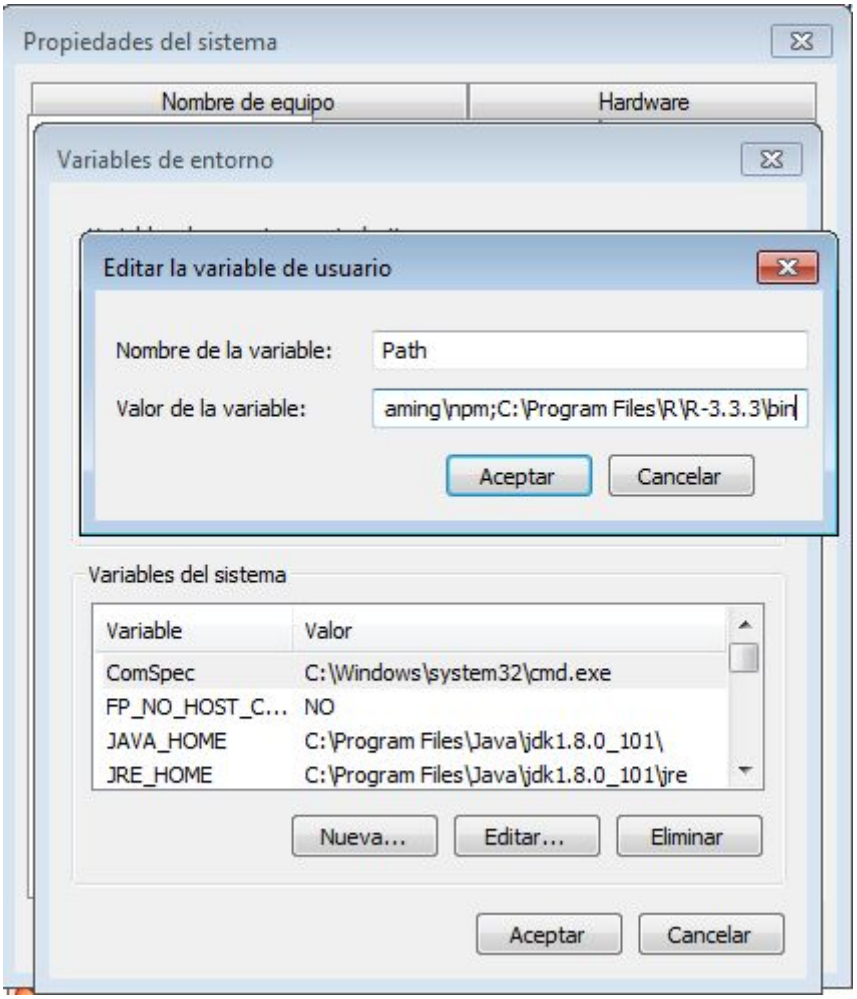
Para instalar RStudio hay que descargar RStudio desde la página oficial de descargas (https://www.rstudio.com/products/rstudio/download/) y seleccionar la versión de escritorio gratuita, tal y como se muestra en la siguiente imagen:



En la instalación para windows, únicamente hay que seguir las instrucciones de instalación que muestra el asistente que se descarga. Una vez finalizada la instalación, ya se puede utilizar RStudio.

R y RScript ya vienen distribuídos como parte de la instalación de RStudio, por loq ue no hay que instalar paquetes o software añadidos.

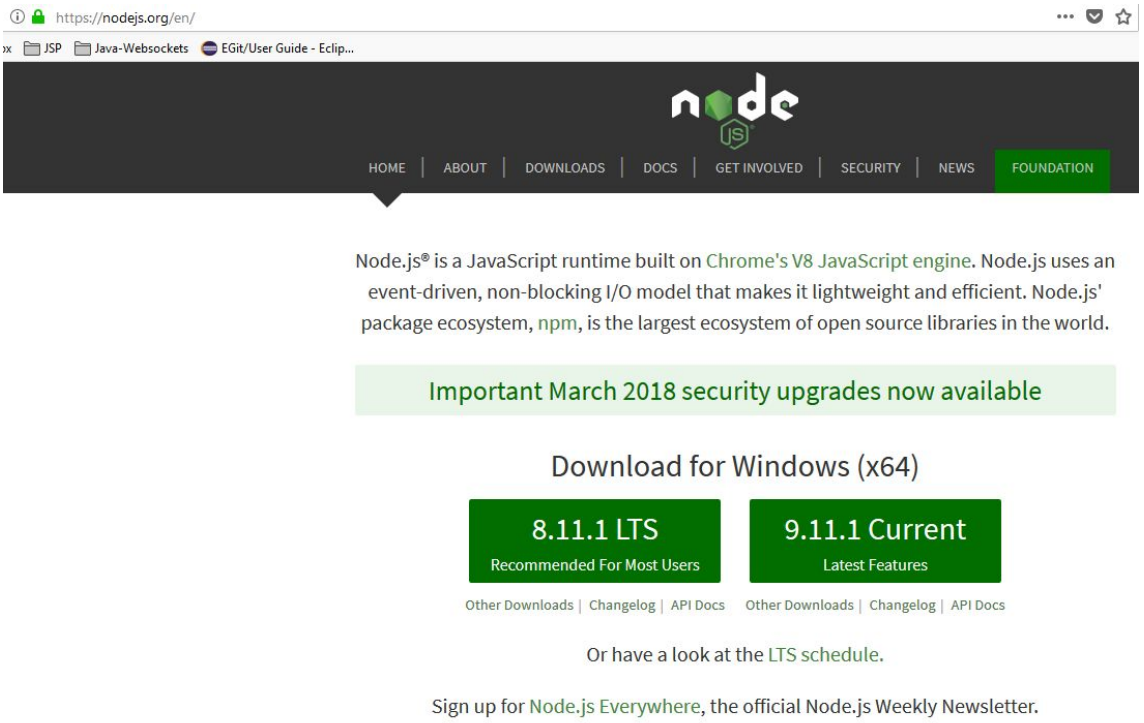
Es un requisito necesario en la instalación añadir a la variable PATH del sistema la ruta a la carpeta donde se encuentra instalado RScript. Si no se añade, no se podrán ejecutar scripts de R desde los procesos automatizados de la plataforma.



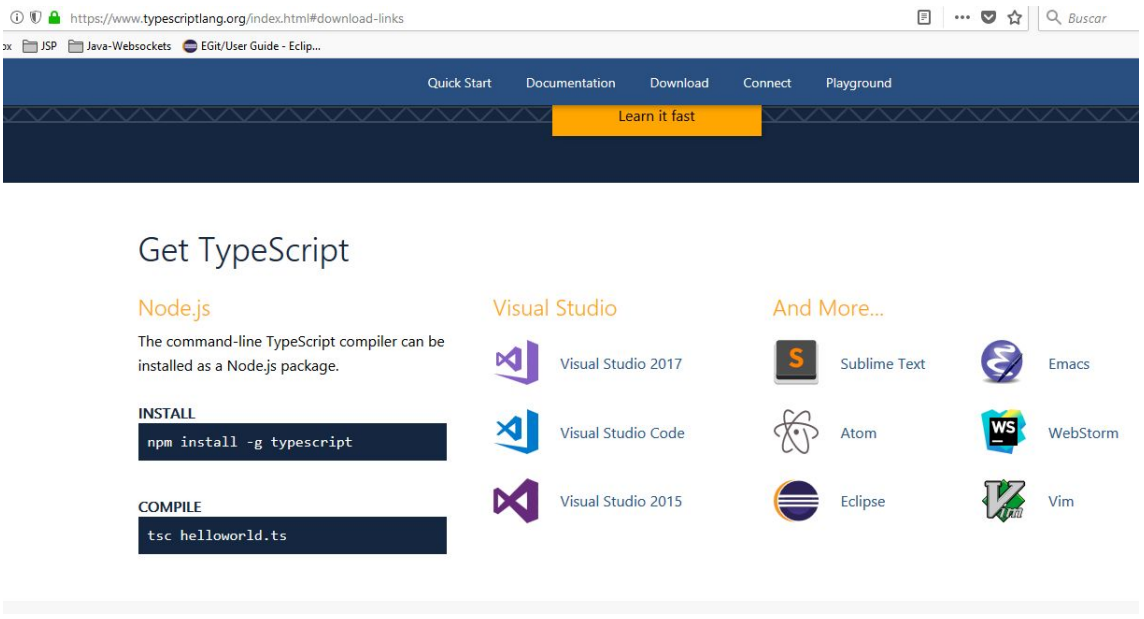
Instalación de NodeJS y Typescript.

NodeJS se utilizará para interpretar el código generado en Javascript en la parte del servidor. Typescript, es un framework de desarrollo cuya finalidad es aproximar el desarrollo javascript al paradigma de la programación orientada a objetos (OOP), ofreciendo tipado para las variables definidas en javascript.

Para instalar NodeJS hay que acceder a su sitio web (https://nodejs.org/en/), descargarlo y seguir el asistente de instalación. La versión utilizada para el desarrollo en este TFM ha sido la 8.11.1 LTS (Long Term Support).



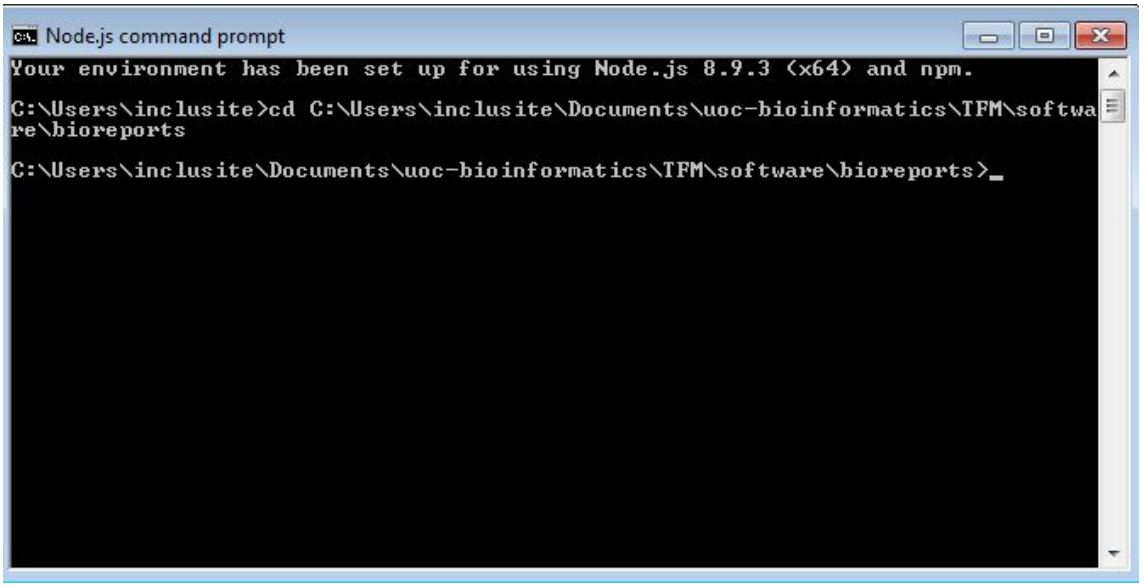
Para instalar Typescript, de nuevo, accedemos a su sitio web (https://www.typescriptlang.org/) y seguimos las instrucciones de instalación, que básicamente consisten en ejecutar un comando del gestor de paquetes de NodeJS previamente instalado.



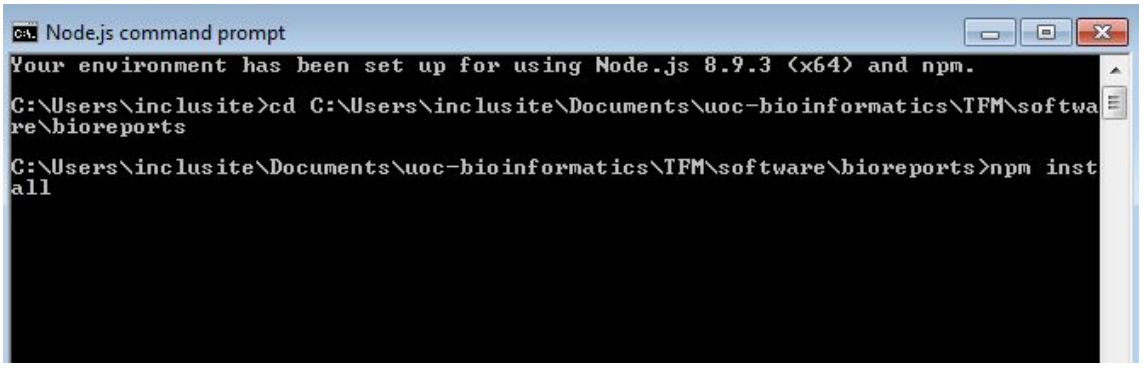
6.2 Anexo II. Puesta en marcha de la plataforma.

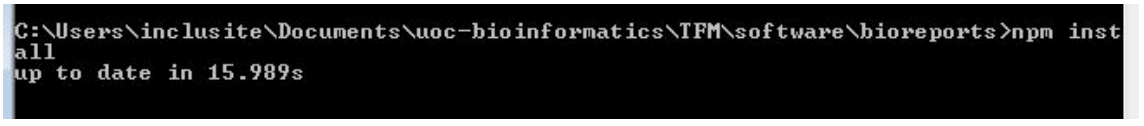
Para poner en marcha el proyecto, son necesarios los siguientes pasos:

1. Abrir una consola NodeJS y posicionarnos en el directorio raíz donde tenemos el código fuente del proyecto.

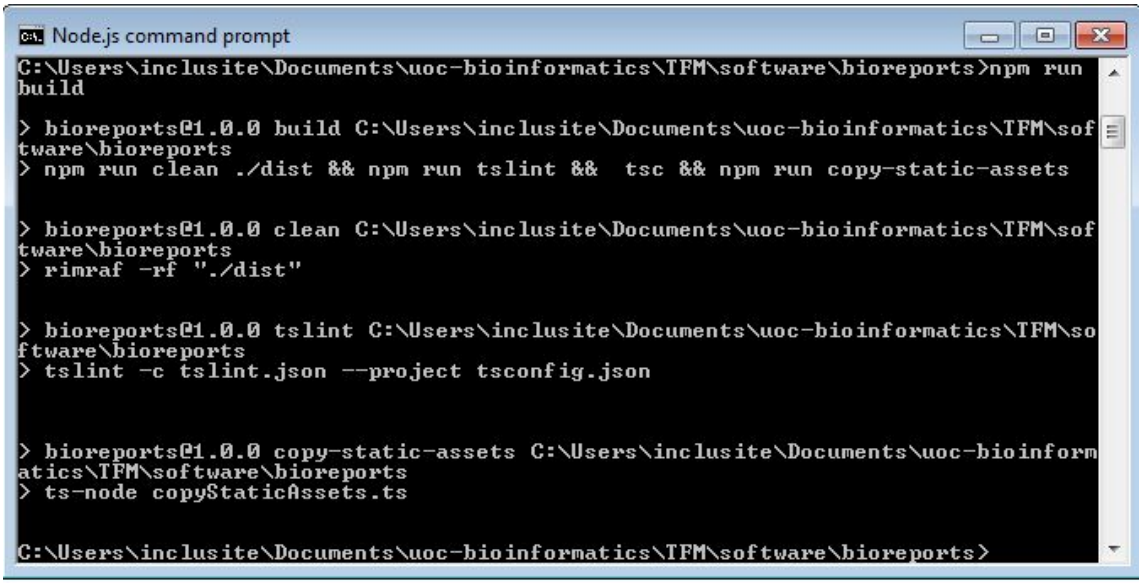


2. Ejecutar la instrucción npm install, que descargará todos los paquetes necesarios.

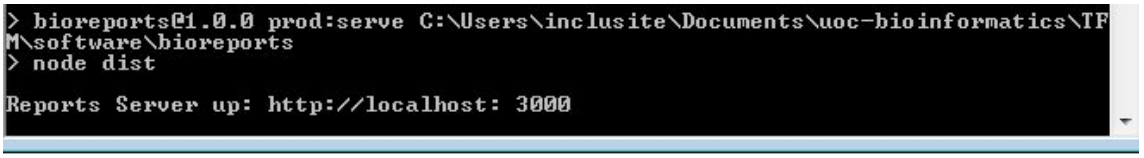




3. Ejecutamos npm run build, para construir el proyecto.



4. Finalmente, podemos arrancar el proyecto con npm run prod:serve



Y si accedemos desde un navegador a http://localhost:3000 veremos la aplicación funcionando correctamente.