Identificación de variantes en regiones codificantes de genes en pacientes colombianos utilizando técnicas de minería de datos

Jennifer Vélez Segura

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ing. Sistemas e Industrial
Bogotá D.C., Colombia
2017

Identificación de variantes en regiones codificantes de genes en pacientes colombianos utilizando técnicas de minería de datos

Jennifer Vélez Segura

Tesis presentada como requisito parcial para optar al título de:
Magister en Bioinformática

Director(a):
Ph.D. Elizabeth León Guzmán

Línea de Investigación:
Minería de datos en Bioinformática
Grupo de Investigación:
MIDAS

Universidad Nacional de Colombia
Facultad Ingeniería, Departamento de Ing. Sistemas e Industrial
Bogotá D.C., Colombia
2017

(Dedicatoria o un lema)

Su uso es opcional y cada autor podrá determinar la distribución del texto en la página, se sugiere esta presentación. En ella el autor dedica su trabajo en forma especial a personas y/o entidades.

Por ejemplo:

A mis padres

o

La preocupación por el hombre y su destino siempre debe ser el interés primordial de todo esfuerzo técnico. Nunca olvides esto entre tus diagramas y ecuaciones.

Albert Einstein

Agradecimientos

Esta sección es opcional, en ella el autor agradece a las personas o instituciones que colaboraron en la realización de la tesis o trabajo de investigación. Si se incluye esta sección, deben aparecer los nombres completos, los cargos y su aporte al documento.

Resumen

El resumen es una presentación abreviada y precisa (la NTC 1486 de 2008 recomienda revisar la norma ISO 214 de 1976). Se debe usar una extensión máxima de 12 renglones. Se recomienda que este resumen sea analítico, es decir, que sea completo, con información cuantitativa y cualitativa, generalmente incluyendo los siguientes aspectos: objetivos, diseño, lugar y circunstancias, pacientes (u objetivo del estudio), intervención, mediciones y principales resultados, y conclusiones. Al final del resumen se deben usar palabras claves tomadas del texto (mínimo 3 y máximo 7 palabras), las cuales permiten la recuperación de la información.

Palabras clave: (máximo 10 palabras, preferiblemente seleccionadas de las listas internacionales que permitan el indizado cruzado).

A continuación se presentan algunos ejemplos de tesauros que se pueden consultar para asignar las palabras clave, según el área temática:

Artes: AAT: Art y Architecture Thesaurus.
Ciencias agropecuarias: 1) Agrovoc: Multilingual Agricultural Thesaurus - F.A.O. y 2)GEMET: General Multilingual Environmental Thesaurus.
Ciencias sociales y humanas: 1) Tesauro de la UNESCO y 2) Population Multilingual Thesaurus.
Ciencia y tecnología: 1) Astronomy Thesaurus Index. 2) Life Sciences Thesaurus, 3) Subject Vocabulary, Chemical Abstracts Service y 4) InterWATER: Tesauro de IRC - Centro Internacional de Agua Potable y Saneamiento.
Tecnologías y ciencias médicas: 1) MeSH: Medical Subject Headings (National Library of Medicine's USA) y 2) DECS: Descriptores en ciencias de la Salud (Biblioteca Regional de Medicina BIREME-OPS).
Multidisciplinarias: 1) LEMB - Listas de Encabezamientos de Materia y 2) LCSH- Library of Congress Subject Headings.

También se pueden encontrar listas de temas y palabras claves, consultando las distintas bases de datos disponibles a través del Portal del Sistema Nacional de Bibliotecas[1], en la sección Recursos bibliográficos.ºpción "Bases de datos".

Abstract

Es el mismo resumen pero traducido al inglés. Se debe usar una extensión máxima de 12 renglones. Al final del Abstract se deben traducir las anteriores palabras claves tomadas del

[1] ver: www.sinab.unal.edu.co

texto (mínimo 3 y máximo 7 palabras), llamadas keywords. Es posible incluir el resumen en otro idioma diferente al español o al inglés, si se considera como importante dentro del tema tratado en la investigación, por ejemplo: un trabajo dedicado a problemas lingüísticos del mandarín seguramente estaría mejor con un resumen en mandarín.

Keywords: palabras clave en inglés(máximo 10 palabras, preferiblemente seleccionadas de las listas internacionales que permitan el indizado cruzado)

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

Motivación

El desarrollo de este trabajo responde a la experiencia de la autora en cuanto a la aplicación de tecnologías de secuenciación masiva aplicadas a la salud de los colombianos, cuyos aportes muestran la relevancia del uso de estas tecnologías en el país que al ser combinadas con técnicas y métodos de análisis de datos a gran escala permitieron realizar un acercamiento de la estructura genética de la población colombiana asociada al fenotipo de los participantes dentro del estudio.

Además de mostrar la importancia de que exista una relación estrecha entre ciencia y tecnología para mejorar el diagnóstico y pronostico de enfermedades presentes en la población aprovechando todas las avances que se encuentran a disposición dentro de nuestro país, generando nuevos conocimientos que tienen un impacto real en la salud.

Estado del arte

En los últimos años con el desarrollo de las tecnologías NGS (Secuencieción de siguiente generación o secuenciación masiva) y otras áreas de la informática se ha introducido una nueva área en las tecnologías de la información conocida como Big Data [1]. En el campo de la bioinformática en concreto es el exoma o secuenciación del genoma completo (WES o WGS), que generan una gran cantidad de información con diferentes aplicaciones en la biotecnología y en la salud de nivel mundial [2]. La enorme cantidad de datos obtenidos por estas nuevas tecnologías presentan son una desafío para ser analizados dado que la estadística tradicional aplicada en genética es poco efectiva para analizar datos de secuenciación de exomas y genomas debido a la gran cantidad de variantes que se obtienen a partir de los experimentos de secuenciación [3, 1].

La aplicación de la secuenciación masiva es posible de aplicar gracias a la reducción de costos y por su capacidad para dar un dar un posible diagnóstico a pacientes que se les sospecha de un síndrome genético de características ambiguas y que con otros estudios no es posible aclarar, o para ser aplicados en paneles genéticos a pacientes que se les sospecha un síndrome especifico [4].

Los datos biológicos en la actualidad están en la escala de petabyte y exabyte,presentando el reto de integrar información y de realizar su posterior análisis, por lo tanto es necesario desarrollar sistemas de información para el manejo y consulta de los datos obtenidos donde los genotipos y los fenotipos,dado que los datos de secuenciación contienen grandes cantidades de información que usualmente se almacena en bases de datos relacionales, después de realizada la anotación de variantes [5] [6].

Estos datos son considerados como "big data" dado que cumplen con los criterios de grandes cantidades de información, velocidad de procesamiento y veracidad de los datos, un ejemplo de esto fue el proyecto de 1000 Genomas, el cual por medio de la secuenciación de genomas completos se genero un sistema de información pública que contiene aproximadamente tres billones de nucleótidos y en el cual la población colombiana no esta correctamente representada dado que se tomo solo una muestra poblacional de la ciudad de Medellín. Además estudios como el perfil de BRCA1 y BRCA2 con la implementación de la secuenciación masiva no tampoco representa la población colombiana[5, 7, 8].

La importancia de la caracterización de la población colombiana esta dada porque la frecuencias de las variantes tienen un alto impacto en la clasificación de la misma siendo las variantes con baja frecuencia poblacional como posibles variantes patogénicas según la ACGM (Asociación Americana de Genética Médica)[9].

Para el manejo de estos tipos de datos se han desarrollado diversas herramientas que incluyen el procesamiento computacional y gestión de estos tipos de datos, así como la creación de buenas prácticas en marco de la integración del análisis de una manera reproducible. Pero el manejo de esta información por parte de los profesionales de las ciencias biolígicos es una gran limitante dado que no tienen fundamentos de programacioín ni conocen los procedimientos que se utilizan normalmente en las ciencias de la computacioń, por lo tanto prefiren utilizar herramientas maś amigables para su uso, pero esto implica un lento procesamiento de los datos ya que los flujos de trabajo que se lleguen a desarrollar son mediante aplicaciones grafícas que consumen maś recursos computacionales [10].

La gestión y análisis de esta información requiere el desarrollo de herramientas que respondan a las necesidades de obtener características relevantes de la información biológica, por ello la implantación de técnicas minería de datos permiten generar hipótesis especificas con respecto a la información genómica [11]. Un ejemplo de esto es la utilización de algoritmos de agrupamiento para encontras grupos de genes que están fuertemente relacionados con estados de evolución de los diferentes estadios en cáncer [5].

En el presente trabajo se muestran la implementación y validación de un pipeline para la identificación de variantes,el diseño e implementación de un sistema para realizar la gestión

de datos para las variantes obtenidas y junto con la información clínica y finalmente un modelo para la minería de datos aplicada en pacientes colombianos.

2. Objetivos

Objetivo General

- Proponer un modelo de minería de datos para la identificación de variantes en regiones codificantes de genes que apoyen el diagnóstico clínico en pacientes colombianos usando ténicas de minería de datos.

Objetivos Específicos

1. Generar un modelo de datos que permita la integración de los datos biológicos de tipo experimental, teórico y clínicos en una muestra poblacional que realice consultas rápidas de los datos almacenados.

2. Diseñar un modelo de minería de datos permita identificar las variantes experimentales que puedan ser patogénicas, teniendo en cuenta parámetros de asociación entre genes e información clínica que posibiliten teorizar posibles predicciones de las variantes.

3. Implementar un modelo de minería de datos que permita identificar posibles variantes patogénicas diferencialmente de las variantes no relevantes y su asociación a la información clńica disponible.

4. Validar el modelo de minería de datos implementado para la identificación de variantes patogénicas en pacientes colombianos utilizando regiones codificantes de genes asociados.

3. Implementación de un pipeline para la identificación de variantes

3.1. Descubrimiento de variantes

La eficiencia del descubrimiento de variantes depende de la exactitud del llamado de las bases (la identificación correcta de cada nucleótido dentro de la secuencia), esto es realizado durante el proceso de secuenciación de alta velocidad con la que se identifican los nucleótidos hace que puedan ocurrir errores en la identificación correcta de las bases, se considera que al momento la exactitud de ese llamado esta alrededor del 99.5 % [12].

Teniendo en cuenta lo anterior es recomendable priorizar la sensibilidad (Buscar tantas variantes como sea posible para evitar perder cualquier variante) sobre la especificidad (Limita la proporción de falsos positivos en un conjunto de variantes) [13].

Existen una serie de pasos para la obtención de variantes la obtención de la calidad de las secuencias y preprocesamiento como la remoción de adaptadores y de nucleótidos con baja calidad (que son erroneamente identificados por el secuenciador),posteriormente sigue el mapeo, post-alineamiento, llamado de variantes, anotación y priorización [14].

Para el presente trabajo se realizaron medición de la calidad de las secuencias y el mapeo, post-alineamiento y el llamado de variantes, siguiendo las buenas practicas para el llamado de variantes [10].

Existen múltiples herramientas para realizar el llamado de variantes tanto de uso privado como open source. Para poder seguir las buenas practicas se hace necesario integrar las diversas herramientas para poder obtener las variantes. Y surge la pregunta de ¿cuáles de todos los métodos y las herramientas son la más apropiada para hacer el llamado de variantes en exones [14][15].

Para dar respuesta a esta pregunta se han seguido la propuesta de buenas practicas para el llamado de variantes propuesto por el Broad Institute que incluyen el procesamiento de los datos,el mapeo (Alinemaiento de las secuencias),descubrimiento de variantes y la recalibración del set de variantes.

Para poder llevar acabo la adecuada implementación se hace necesario la utilización de HPC (Computación de alto desempeño) donde la utilización de un clúster para bioinformática presentan una gran apoyo para el procesamiento y análisis de datos incluso es un requisito de algunos módulos para poder implementarse adecuadamente [10].

Los datos que fueron procesados en el presente trabajo son secuencias de 4813 exones humanos se obtuvieron de kit de Illumina TruSight One en muestras de sangre periférica. Estos datos fueron donados por el Centro de Investigaciones en Genética Humana y Reproductiva GENETIX S.A.S dirigido por la Dra Claudia Serrano Médico Genetista. También se realizo una obtención de variantes a partir de un exoma completo público de la muestra NA12878,los datos fueron obtenidos via ftp en la siguiente direcciones:

`https://s3.amazonaws.com/bcbio_nextgen/NA12878-NGv3-LAB1360-A_1.fastq.gz`

`https://s3.amazonaws.com/bcbio_nextgen/NA12878-NGv3-LAB1360-A_2.fastq.gz`

Y el archivo bed para filtrar las variantes que se encuentran dentro del genoma completo de la muestra se obtuvo de la siguiente pagina para el genoma hg19:

`ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/analysis/`

`Illumina_PlatinumGenomes _NA12877_NA12878_09162015/hg19/8.0.1/NA12878/`

Se utilizo el modulo de omics-pipe que presenta el siguiente pipeline que es acorde con las buenas practicas para el llamado de variantes:

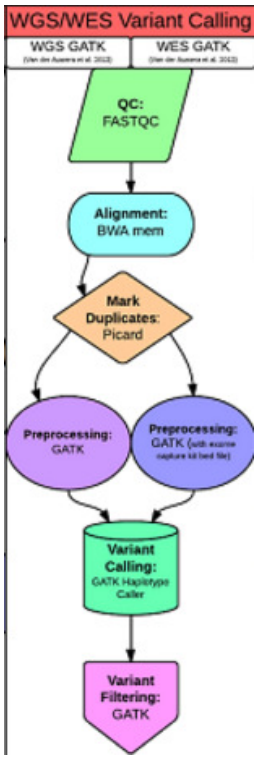


Figura 3.1.: Pipeline basado en las buenas practicas para el llamado de variantes.

3.1.1. Implementación en de herramientas dentro del clúster

Las herramientas bioinformáticas seleccionadas se implementaron el clúster de la Universidad de los Andes que cuenta con las siguientes características:

⇒ Un nodo mastes con 2 procesadores Intel Xeon E5-2695 – 24 cores 48 con HT / 192 GB RAM (230Gflops), 300 GB de Disco duro.

⇒ Se tienen 19 nodos de trabajo con 2 procecesadores Intel Xeon E5-2695 – 24 cores 48 con HT / 192 GB RAM (4.378Tflops), 300 GB de de Disco duro.

⇒ Se cuentan con otros 7 nodos de trabajo con 4 procesadores AMD Opteron 6282 SE – 64 cores / 128 GB RAM (3.659Tflops), 200 GB de Disco duro.

⇒ 1 GPU tesla K20 como nodo de trabajo con 2 Procesores Intel Xeon X5690 - 12 cores / 192 GB RAM (3.659Tflops), 1.6 TB de Disco duro.

Y se instalo el modulo para python de omics-pipe, para python 2 con la herramienta de R y las librerías que solicita omics-pipe[10] , el algoritmo BWA, samtools,vcftools, GATK 3.5,picard, FASTQC y pbs-drmma.Una vez instalados los programas se procesaron las muestras dentro del clúster.

3.2. Resultados

3.2.1. Reporte FASTQC

Este reporte utilizando la herramienta FASTQC presenta inicialmente un resumen del estado de las secuencias obtenidas, ya que toma el archivo fastq y lee las métricas de calidad de cada una de las bases y genera un reporte general en formato HTML. Ya que es interactivo y genera varios módulos [16].

Este reporte no presenta fallas dentro del análisis. A continuación se muestra un el primer modulo del reporte FASTQ obtenido de un dato experimental de una secuenciación de 4813 genes que resume el estado general de las lecturas obtenidas para este caso, la figura 3.2 muestra que la calidad de las secuencias es mayor a 30, el reporte general también muestra que no hay secuencias adaptadoras, que presenta una distribución media del largo de las secuencias aceptable y que no hay secuencias sobrerpresentadas.

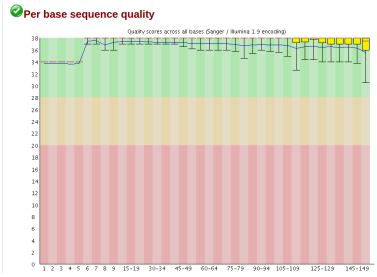


Figura 3.2.: Calidad del llamado de bases en una secuencia Estadísticas básicas del reporte FASTQ

Para el caso de esta muestra la calidad es óptima en todos los datos obtenidos y no requieren de ningún tipo de trimming ya que la mayoría de las posiciones dentro de la secuencia se encuentran por encima del un valor por encima de 34 y el cual el valor mínimo es 20 (este valor representa el (Q_{PHARED}) que implementa el secuenciador [16].

3.2.2. Resultados de variantes de Illumina vs Variantes de Omics

Inicialmente se obtuvieron 63515 variantes una vez que se ejecuto el pipeline de omics para la obtención de variantes, siguiendo los protocolos de buenas practicas y los protocolos de GATK quienes recomiendan generar variantes altamente sensibles y poco precisas, esto con el fin de no perder variantes que se encuentren dentro de las secuencias obtenidas, por ello se muestra una gran cantidad de variantes que no corresponden con las variantes verdaderas [13].Dentro del pipeline solo se encuentra el proceso de llamado de variantes y no el proceso

de filtrado de las mismas y que debío ser implementado de manera manual.

A partir de la aplicación del pipeline se obtuvieron los siguientes resultados representado en tabla (**3-1**):

	Variantes			
	SNP	Indels	Desconocida	Total
Variantes Omics	54538	8855	122	63515
Variantes Calibradas	10425	828	44	11297
Variantes Illumina	9601	436	28	10065

Tabla 3-1.: Tabla de Variantes obtenidas.

De las variantes sin hard filtering se obtuvieron 54538 SNP, Indels 8855 y 122 variantes desconocidas, que se representan el siguiente gráfico 3.3:

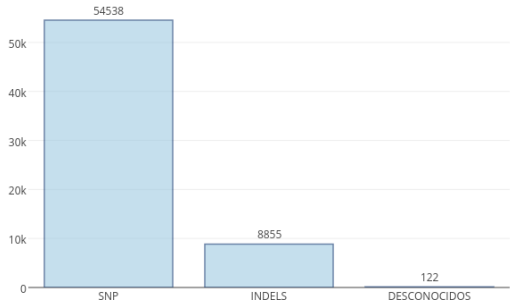
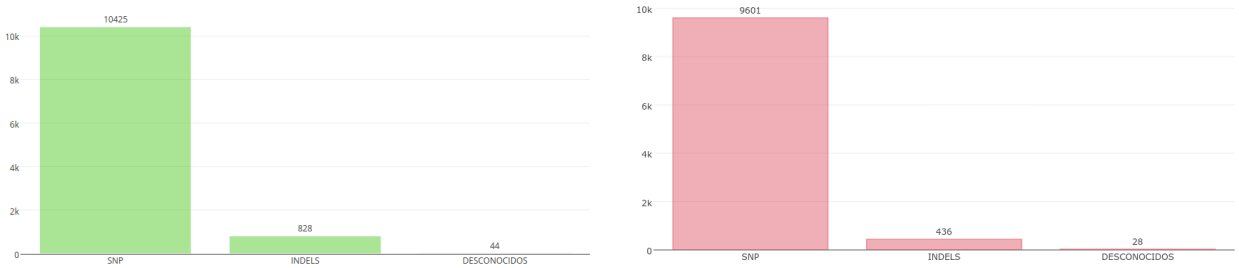


Figura 3.3.: Variantes obtenidas por Omics pipe

Una vez realizado el hard filtering se obtuvo los siguientes resultados: 10425 SNP, 828 Indels y 44 desconocidos, también se tiene las variantes reportadas para el mismo individuo desde la plataforma de illumina con los siguientes resultados: 9601 variantes, 436 indels y 28 desconocidas y respresentado por la figura 3.4:



(a) Variantes obtenidas por Omics pipe (b) Variantes calibradas

Figura 3.4.: Variaciones de la muestra

Se realizo un distribución de las variantes según cada técnica sin filtrado (para el caso de omics) para el siguiente gráfico mostrados en las siguientes figuras 3.5 y 3.6 :

Cromosoma	Variantes Omics	Variantes Calibrad	Variantes Illumin
1	5600	1000	964
2	4675	782	701
3	3546	676	540
4	2959	481	414
5	2702	502	438
6	3384	556	795
7	3313	502	447
8	2306	415	372
9	2799	491	401
10	2998	524	381
11	3289	776	610
12	3192	607	535
13	1433	223	191
14	1530	275	258
15	2233	400	327
16	2529	478	472
17	3172	681	611
18	1434	184	184
19	2651	704	544
20	1245	244	195
21	1135	178	136
22	1315	271	262
X	1615	299	249
Y	437	4	10

Figura 3.5.: Distribución de variantes a lo largo de los cromosomas

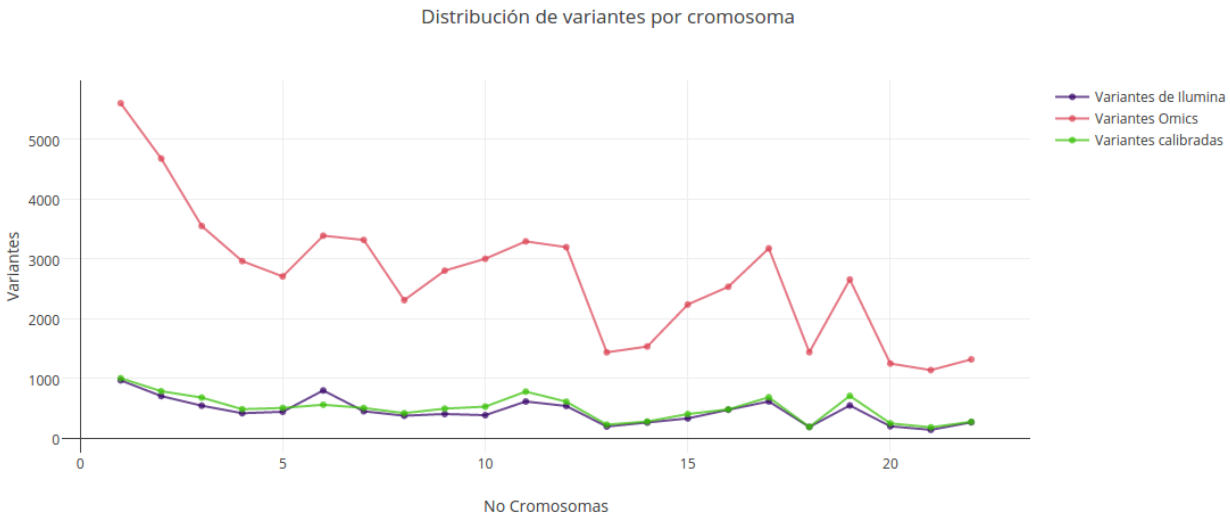


Figura 3.6.: Distribución de variantes a lo largo de los cromosomas

Se observa la distribución de las variantes a lo largo del genoma, inicialmente las variantes obtenidas son en grandes cantidades para el modulo de omics, pero conservan el patrón de distribución es similar para los tres casos,incluso cundo se realiza el hard filterin las diferencias en cuanto a la distribución de las variaciones es similar, siendo la mayor para el cromosoma 1 y la menor para el cromosoma Y.

Al realizar la comparación entre los dos archivos vcf se obtuvieron los siguientes resultados los archivos vcf de Illumina y los de Omics comparten 49.4 % d y 44.0 % de las variantes, y difieren entre un 50.6 % para Illumina y 56-0 % para los datos de omics pipe. Como se refleja en el siguiente diagrama (3.7):

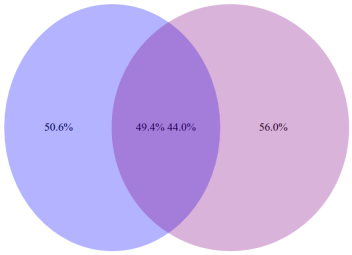


Figura 3.7.: Diagrama de relación entre variantes comunes de Omics y de Illumina

3.3. Variantes de Exoma vs Variantes de Omics

Para la validación del pipilene se corrio un exoma público de NA12878-NGv3-LAB1360 que pertenece a una mujer que tiene una variación en el gen CYP2C19 donde tiene una transición de una Guanina por una Adenina en la posición 681 del exón 5, que causa un cambio en el marco de lectura del ARNm a partir del aminoácido 215 y produce un códon de parada prematuro en 20 aminoácidos corriente abajo produciendo una proteína no funcional (*Información obtenida de Coriell Institute for medical reseach*). Se descargo el archivo bed para filtrar las variantes que se encuentran dentro del genoma completo de la muestra se obtuvo de la siguiente pagina para el genoma hg19.

Una vez obtenidas las regiones se realizo el proceso de hard filtering para el vcf obtenido por el pipe de omics y por el generado por vcftools teniendo los siguientes resultados mostrados por la tabla **3-2**:

	Variantes Exoma			
	SNP	Indels	Desconocida	Total
Variantes Omics	30893	3324	0	34217
Variantes Públicas	29749	3101	0	32850

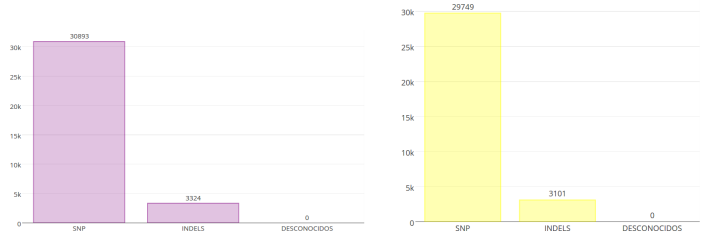
Tabla 3-2.: Tabla de Variantes obtenidas a partir de un exoma.

Donde se observa una diferencia de 1367 en el total de las variantes encontradas, para los SNPs se encuentra una diferencia de 1144 y 223 para los indels, no se encuentran variantes que no hayan sido correctamente identificadas. Presentado en los siguientes gráficos 3.9

La distribución de las variantes a lo largo de los cromosomas se presenta en la siguiente tabla 3.8:

Cromosoma	Variantes publicas	Variantes Omics
1	3329	3438
2	2483	2586
3	1887	1906
4	1497	1546
5	1358	1377
6	1489	1540
7	1581	1658
8	1088	1111
9	1514	1595
10	1406	1529
11	2183	2256
12	1669	1698
13	720	736
14	1044	1111
15	1192	1230
16	1294	1372
17	1781	1921
18	545	607
19	2001	2060
20	807	820
21	490	500
22	755	795
X	737	755

Figura 3.8.: Distribución de variantes a lo largo de los cromosomas para los exomas.



(a) Variantes obtenidas por Omics (b) Variantes públicas

Figura 3.9.: Variaciones de la muestras dentro del exoma

Y la representación gráfica de las variantes sobre la distribución a lo largo de los cromosomas figura 3.10:

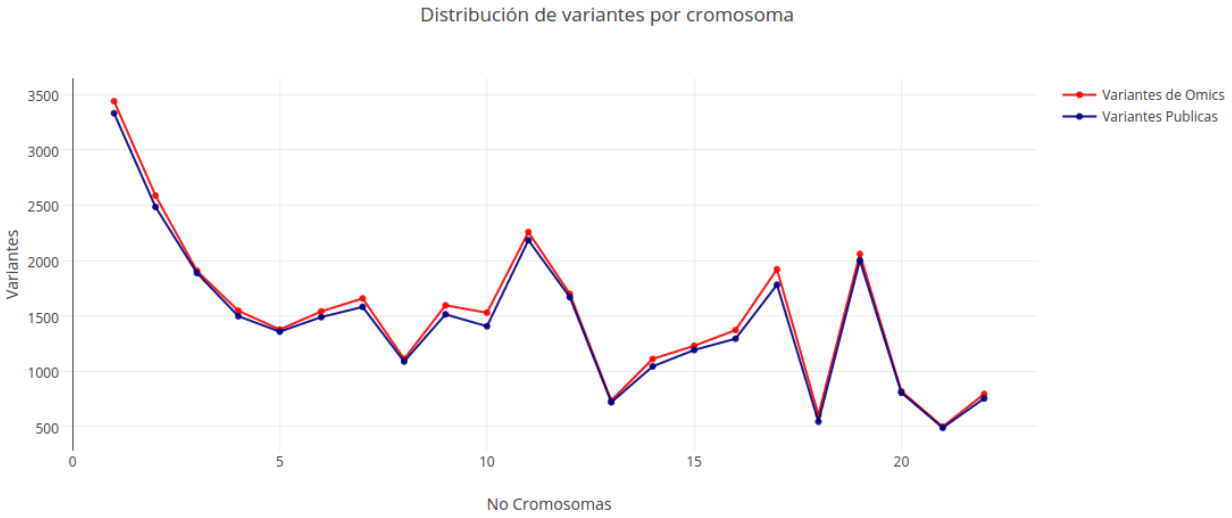


Figura 3.10.: Distribución de variantes a lo largo de los cromosomas para los exomas

En la figura 3.10 se observa el comportamiento de la distribución de las variantes para los datos públicos y los datos obtenidos para el pipeline donde se encuentran un comportamiento similar de la distribución, pero se observa que aún hay una mayor cantidad de variantes obtenidas por el pipeline. En la siguiente figura se observa el comportamiento de las variantes públicas con respecto a las variantes del pipeline.

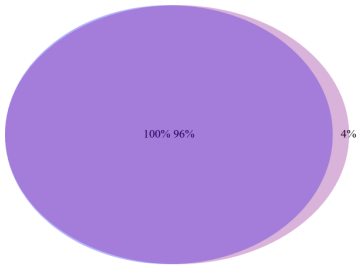


Figura 3.11.: Diagrama de relación entre las variantes publicas y las obtenidas por el pipeline

El diagrama de la figura 3.11 muestra la comparación de las variantes obtenidas y su respectiva concordancia done el 100 % del exoma esta representado en las variaciones encontradas mienteras que un 96 % de las variaciones obtenidas por omicspipe son un 96 % dejando solo un 4 % de las variantes no encontradas dentro del exoma publico.

GATK realiza un reporte de la evaluación cunado se comparan dos archivos de distintas variaciones, este puede ser abierto como un archivo de texto o cargado directamente en R utilizando la librería *gsalib* quien lee el archivo *merged.eval.gatkreport*, esto genera una lista que tiene anidados varios data.frame, dentro de ellos para este caso se tomo el *Validation-Report)* que genera una tabla con los falsos positivos, falsos negativos, calcula la sensibilidad

y la especificidad y el valor predictivo positivo (PPV). Que para nuestro caso de nuestro exoma se encuentran en la tabla **3-3**:

TP	**FP**	**FN**	TN
32110	0	1033	0

Tabla 3-3.: Tabla de validación 1.

La tabla **3-3** refleja que para el conjunto de datos no hay falsos positivos ni verdaderos negativos, pero si falsos negativos, es decir 1033 variantes del conjunto de datos obtenidos. (*GATK para calcular estas métricas compara contra una base de datos que el usuario disponga para determinar variantes existentes*). La tabla **3-4** muestra la sensibilidad, especificidad y el valor predictivo positivo (PPV).

Sensibilidad	**Especificidad**	**PPV**
96.88	100	100

Tabla 3-4.: Tabla de validación 2.

Donde se tiene que existe una sensibilidad de 96.88 %, una especificidad del 100 % y un PPV de 100.Además después de la limpieza de los datos se realizo la anotación del archivo vcf obtenido, para el gen CYP2C19 utilizando la versión gráfica de annovar [17] y obteniéndose el siguiente resultado:

```
chr10,96541616,96541616,G,A,exonic,CYP2C19,synonymous SNV, CYP2C19:
NM_000769:exon5:c.G681A:p.P227P
```

La representación escrita informa el cromosoma, la posición dentro del genoma y el cambio de posición en el genoma, las siguiente es el cambio Guanina por Citocina (representado por sus letras) tipo de variación que en este caso es sinónima, el nombre del gen, su identificador, ubicación exonica y cambio en la posición del exón, finalmente se tiene el cambio en la proteína (No sigue exactamente la nomenclatura de HGVS), esta variación se confirmo también realizando la visualización por medio de la herramienta IGV conectado al clúster.

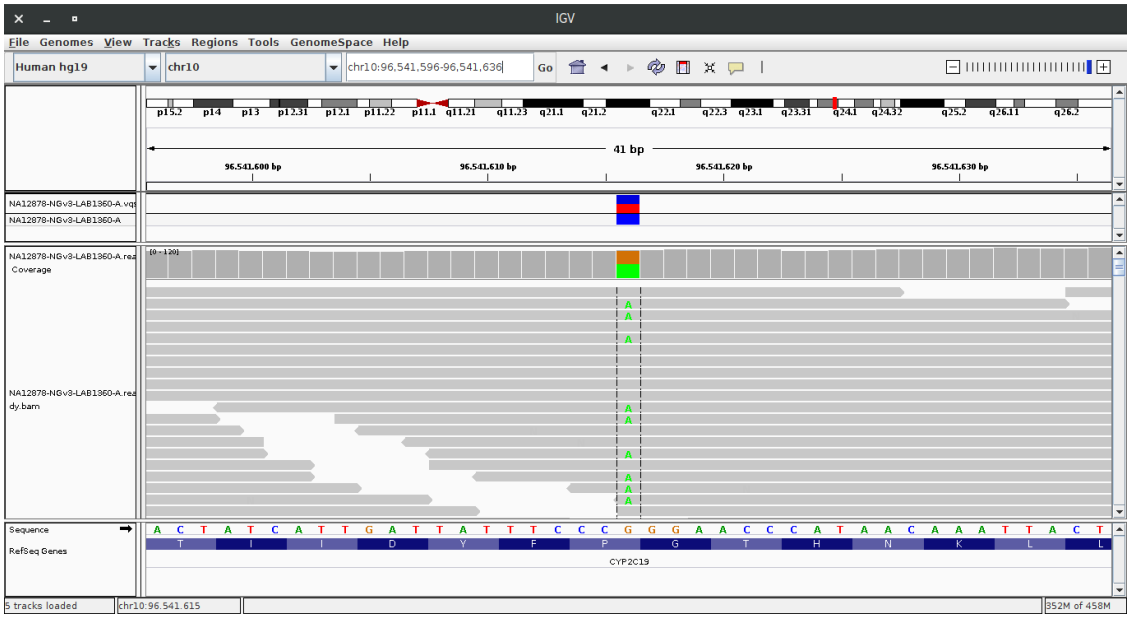


Figura 3.12.: Imagen de la variante presenten en el exoma público

3.4. Discusión

3.4.1. Preprocesamiento

La revisión de las metricas dadas por el FASTQC report muestran el estado de como están las secuencias antes de ser procesadas, aunque a nivel experimental no dependiendo de las condiciones y el tipo de muestra los niveles de calidad terminan bajando de manera sustancial y depende del analista tomar la decisión de remover secuencias o mantenerlas ya que los diferentes módulos presentan diversas meticas de evaluación de las secuencias [16].

El presente conjunto de secuencias FASTQ se encuentra con buenos parametros de calidad, aunque algunos modulos presentan falla, el percentil, el porcentaje de GC, la distribución del largo de las secuencias, los niveles de duplicación de las secuencias y los valores de K-mer y las secuencias en secuencias cortas de 7 nucleótidos, representan que dentro del conjunto de datos estas secuencias cortas están en la parte inicial de la mayoria de las lecturas obtenidas en la muestra y que posiblemente son secuencias duplicadas que no pertenecen al conjunto de secuencias real, apesar de que no se encuentran adaptadores, ni representaciones al final de las lecturas. Esto puede llevar a dos caminos, el primero que estas secuencias sean parte de un adaptador (llama la atención que no se encuentren al final de la secuencia) o que sean errores propios del proceso de secuenciación durante la hibridación de las secuencias y sean representados como duplicaciones de las secuencias originales [16][18].

Además existen otras características que pueden generar impactos negativos dentro del analisis de datos de NGS divididas en dos grupos [19]:

1. Lecturas con baja calidad: Las calidades de las lecturas generadas por un secuenciador pueden degradarse durante el proceso de corrido y es común ver fallas al final de la lectura o tener secuencias duplicadas a partir de la amplificación por PCR durante la construcción de las librerías [19].

2. Contaminación de las lecturas de especies conocidas o no conocidas en la secuencia objetivo, este error es frecuente y puede se causado por un experimento artificial durante la preparación de la muestra, la construcción de la libreria o otro paso experimental, sin embargo las muestras de ADN pueden contener algunos nucleótidos de otras especies, las cuales son difíciles de excluir de manera experimenta y por lo tanto si se cree que hay una contaminación lo ideal es realizar un triming de las secuencias para remover la contaminación. **Nota:** Siempre y cuando estén en una baja proporción [19].

Las secuencias que se observan pueden ser duplicados de PCR que son un problema critico cuando los fragmentos están sobre amplificados durante la preparación de las librerías, estos duplicados pueden aumentar a frecuencia alelica e incluir una detección erronea de variantes, esto es muy común los datos de metagenomica, pero en nuestro caso los datos no son datos de metaegenomica si no de un solo individuo llama la atención de que solo esten al inicio de las lecturas y que el final de las lecturas este adecuado esto podría indicar que más que un duplicado de PCR pueda ser un error de secuenciación al inicio de cada nuevo ciclo.[20].

Teniendo en cuenta lo anterior se puede inferir que las secuencias duplicadas son bajas y que la calidad de los datos obtenidos son adecuados para continuar con el procesamiento de las secuencias FASTQ, dentro del pipeline se cuenta con una herramienta para remover las secuencias duplicadas (PICARD) y así obtener una calidad optima de los datos.

3.4.2. Variantes obtenidas

Variantes de illumina y omics pipeline

En los datos obtenidos para illumina inicialmente reflejados en la tabla **3-1**, muestran una alta discordancia ya que inicialmente las variantes no se les aplicó un segundo filtro, siguiendo las recomendaciones de GATK , donde por el pipeline de Omics tiene por defecto el variant quality score racalibration (VQRS) que se basa en machine learning para filtrar las variantes y generar una alta sensibilidad verdadera, que es el método más recomendado, pero tiene limitaciones estadísticas y es más robusto que el hard filtering, este es recomendado para datos pequeños [13].

Al realizar una calibración de los datos con la calidad y con hard filtering en GATK se obtiene una similitud entre la cantidad de variantes obtenidas por omics con respecto a Illumina, pero aún es posible ver que la distribución de las variantes es similar para ambos conjuntos de datos (véase la figura 3.6) y se acerca más después de realizar el filtrado. Esto se presenta debido a que no existe una formula para determinar cuales anotaciones y filtros son adecuados, además el VQSR genera datos de entrenamiento para determinar las variaciones y se hacen recomendaciones según lo que se ha observado empíricamente dentro del desarrollo de los algoritmos [13].

A pesar de que la distribución de las variante es similar, aun con el filtrado de las variantes existe que la concordancia entre ambas técnicas tiende a ser del 50 % (véase la figura 3.7), aunque illumina utiliza GATK la versión implementada es la 1.6 que en este momento no cuenta con documentación (https://www.broadinstitute.org/gatk/guide/version-history) que illumina utiliza la versión 1.6 y la función UnifiedGenotyper que presenta algunas inconsistencias para la identificación de indels, mientras que la versión de GATK 3.5 utiliza la función HaplotyperCaller que mejora el llamado de variantes, y corrige algunas inconsistencias para la identificación de indels [21]. Además es la función recomendada para organismos diploides, este se enfoca en dos tipos de identificación inicialmente los SNPs y los indels, y puede identificar cuando hay varios tipos de variantes cercanas a otras [13].

Illumina no provee los parametros utilizados para hacer el llamado de variantes lo que dificulta la comparación entre este pipeline y las variantes reportadas por illumina, además el formato del VCF es el 4.1 y en la mayoria de las variantes no reporta el valor de la Qual (calidad) para hacer un filtro con el archivo aunque para GATK los valores para el llamado de variantes no son modificados de manera significativa si se realiza un filtro de este tipo [2]. Además de que la combinación de BWA con HaplotypeCaller, presentan una mejora con respeto a la identificación de SNPs (BWA-men) y HaplotypeCaller para la identificación de indels [15].

Variantes con un exoma NA12878.

Para este estudio se utilizo una muestra del genoma completo de la muestra NA12878 son de 34,886 variaciones [15] en el presente estudio 32850 y el pipeline obtuvo un total de 34217, lo que permite inferir que las variante identificadas son solo de 2036 variaciones (dependiendo de las muestras y los genes que fueron secuenciados) y que se realizo un muestreo partir de un archivo bed. Además si se aplica un filtro para retirar las variaciones con baja calidad, el llamado de variantes de GATK mejora de manera significativa si necesidad de hacer cambios en el preprocesamiento de los datos [22]

Las dos resultados presentan una distribución similar en cuento a las variantes por cromoso-

ma y no hay variantes desconocidas dentro de la muestra,esto se debe a la alta curación que tiene este exoma, la figura 3.10 presenta la distribución a lo largo de los cromosomas donde se presenta leves diferencias entre los datos públicos y los datos generados por el pipeline con una diferencia del 4 % entre las dos resultados, no existen falsos positivos ni verdaderos negativos identificados dentro del conjunto de los datos del pipeline, se presenta una sensibilidad del 96 % que es alta , dado que las calibraciones y los algoritmos presentan falencias reales para la identificación de variantes [13]. Esto se puede corregir por dos vias, aplicando un filtro de Quality by Depth (QD) ¿= 4 and Fisher Strand Bias (FS) =¡30 para dar un balance a la sencibilidad y la especificidad [23] o aplicando múltiples pipelines.

La no existencia de falsos negativos y verdaderos positivos, esta supeditada al hecho de que se tomo una muestra de las muestras comunes, es decir que tanto en la muestra a comparar con la obtenida se van a ver las posiciones entre los datos analizados, aún con esta limitante acerca de la posición de las variantes en la región génomica se logra ver el error que se esta obteniendo dentro del pipeline, aunque la precisión es alta y no hay false descovery rate (FDR).

La sensibilidad de un solo pipeline esta en promedio de 95 % al 99 % ,que esta dentro del rango de aceptabilidad para la identificación de las variantes [24]. Para nuestro pipeline tenemos una precisión de 100 %. Lo que muestra que hay baja probabilidad de error.

Al realizar la anotación se logro encontrar una de las variantes reportadas para el exoma, en el gen CYP2C19 en la misma posición reportada, con la misma variación mostrando la concordancia entre los resultados del pipeline y la muestra original.

Para ambos estudios se presentan archivos intermedios de gran tamaño como son los bam y bai que permiten la visualización de las variantes que pesan entre 6 y 15 gigas para un exoma completo, los datos iniciales pueden pesar entre 1 y 3 gigas (fastq) dependiendo de la cantidad de genes que se hallan secuenciado, lo que requiere de la disponibilidad de un computo para su almacenamiento y procesamiento.

3.5. Conclusiones

La validación de un pipeline para la identificación de variantes requiere la utilización de herramientas computacionales de HPC para hacerse de manera eficiente. Es necesario que se tengan conocimientos de programación básica y biología molecular, con el fin de definir los para metros óptimos para la implementación un pipeline.

La cantidad de herramientas y parámetros para aplicar son diversos y dependen del investigador decidir cuales son los mejores y que filtros van a ser utilizados, dado que a pesar de la existencia de protocolos no hay un consenso de cual o cuales son los mejores y estos

dependen del conjunto de datos obtenido.

El llamado de variantes es bueno para el presente estudio, pero hay la posibilidad de mejorar la implementación de los parámetros de filtrado y el proceso de anotación (implicación del cambio de las variantes), además generar un pipeline alternativo para la verificación de las variantes que están siendo identificadas y poder aumentar la sensibilidad.

Es necesario crear o generar la manera de optimizar los tiempos de ejecución de las tareas, de una manera más eficiente a la dada por el omics pipe.

4. Sistema de información

4.1. Introducción

Las nuevas tecnologías de análisis genético son fáciles y económicas de hacer lo que genera una gran cantidad de datos biológicos y lo que hace que los biólogos trabajen cada vez más con las nuevas tecnologías de análisis genético, haciendo que los biólogos trabajen más y más computacionalmente.Especialmente mediante el uso de tecnologías de secuenciación (NGS) y presentar un reto para integrar y almacenar la información, pasos que son necesarios para su posterior análisis [5, 25].

La gran cantidad de datos presentan un reto para organizar y manejar datos que crecen de manera exponencial y que son de diversos tipos, dado que los datos son generados a diferentes niveles y con diferentes métodos (ejemplo: Variantes de exones o imágenes de patología), datos que a su vez deben ser almacenados en distintas formas, esta situación muestra una seria dificultad para realizar un análisis integral de los datos [25, 5].

El mayor de los retos es crear herramientas que permitan al investigador acceder a la información fácilmente y que pueda tener una base de datos intercalable,donde pueda consultar, analizar y actualizar la información de sus experimentos [5].

En el campo clínico esto representa un reto aun mayor dado que se hace necesario recolectar los datos genéticos junto con los datos clínicos para poder hacer análisis más acertados y a gran escala [26].

El problema de la heterogeneidad de los datos se aplica igualmente a los datos clínicos que describen pacientes individuales y además a los datos biológicos que caracterizan nuestro genoma. Específicamente, las bases de datos son altamente heterogéneas con respecto a los modelos de datos que emplean, los esquemas de datos que especifican, los lenguajes de consulta que soportan y las terminologías que reconocen [27].

Por ello se hace necesario que se utilicen herramientas para la gestión de la información, por ejemplo django que es un web framework de alto nivel desarrollado en python fomenta el desarrollo rápido y limpio, para la creación de aplicaciones web, es de código abierto y gratuito.Se basa en los principios de desarrollo rápido, manejo de la seguridad y es altamente escalable .

Dentro de las muchas aplicaciones que tiene django una es el manejo y gestión de bases de datos a través de los módulos de python. Ver documentación https://www.djangoproject.com/ .

4.2. Metodología

Nosotros a continuación proponemos la utilización de una base de datos con la información clínica disponible y las anotaciones de variantes obtenidas por medio de un pipeline validado previamente para la anotación de variantes, basado en la propuesta de Fisch y colaboradores en 2014 [10] y anotados con annovar web service [17]. Las historias clinicas fueron transcritas manualmente y cargadas desde un archivo de texto plano con un formato especifico.

Se genero una base de datos utilizando Django 1.10 en python 3 y la librería grapelli, que se conectan a MySQL 5.7.17 desde un PC portatil HP-pavilion 360 con ubuntu 16.04 LTS. Django genera los modelos de EER pero permite su modificación para optimizar la velocidad de las consultas.

4.3. Resultados

Los resultados obtenidos fueron una aplicación con una interfaz que permite a los usuarios con poco conocimiento de programación analizar los datos de variantes y su resumen de la historia clínica.

Figura 4.1.: Interfaz de ingreso para administrar la base de datos.

Inicialmente la figura (4.1), muestra la solicitud de usuario y contraseña para acceder a la aplicación, es diferente a la base de MySQL que puede tener una contraseña igual o diferente a la interfaz gráfica .

Figura 4.2.: Interfaz de administración.

La figura (4.2), muestra el sitio de administración donde se encuentran los usuarios permitidos, las bases de datos a consultar y muestra un histórico de las actividades recientes.

Desde esta interfaz se puede agregar un grupo, más usuarios y pacientes y/o variantes dando click en el signo más sin necesidad de hacer la carga directa a MySQL ya que Django se encarga de hacer la carga.

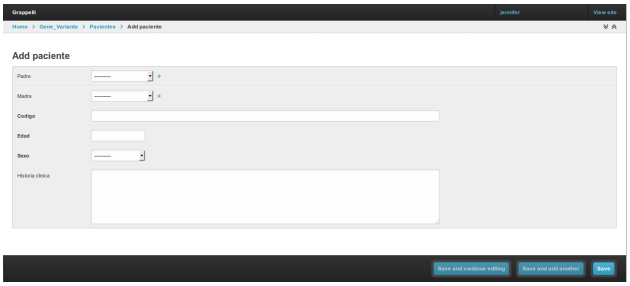


Figura 4.3.: Ingreso de pacientes.

En la figura (4.3) se muestra el formulario para ingresar una nueva historia o de modificar una historia clínica de un paciente de manera manual.

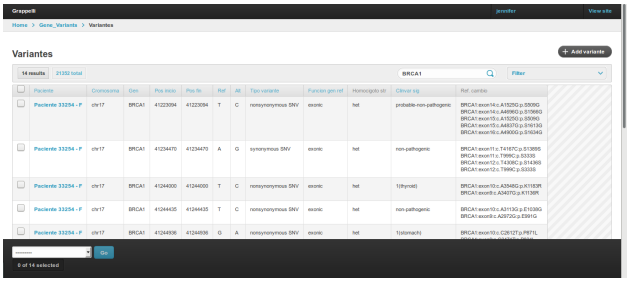


Figura 4.4.: Consulta a variantes

La figura (4.4) muestra una consulta de las variantes que se tienen cargadas en la base de datos para el gen BRCA1, donde nos muestra una consulta de las variantes con su anotación filtrada mediante un script de python antes de cargar las anotaciones de la tabla obtenida por annovar para cada paciente. Desde esta misma interfaz se puede hacer consultas de pa-

cientes que se deben eliminar, en la parte inferior se encuentra la opción.

Si se desea hacer modificaciones a los datos del paciente también es posible hacerlo desde esta misma interfaz seleccionando el código del paciente, que lleva a la tabla de genes_varante_paciente que contiene el formulario de la historia clínica con los datos cargados para ser modificados.

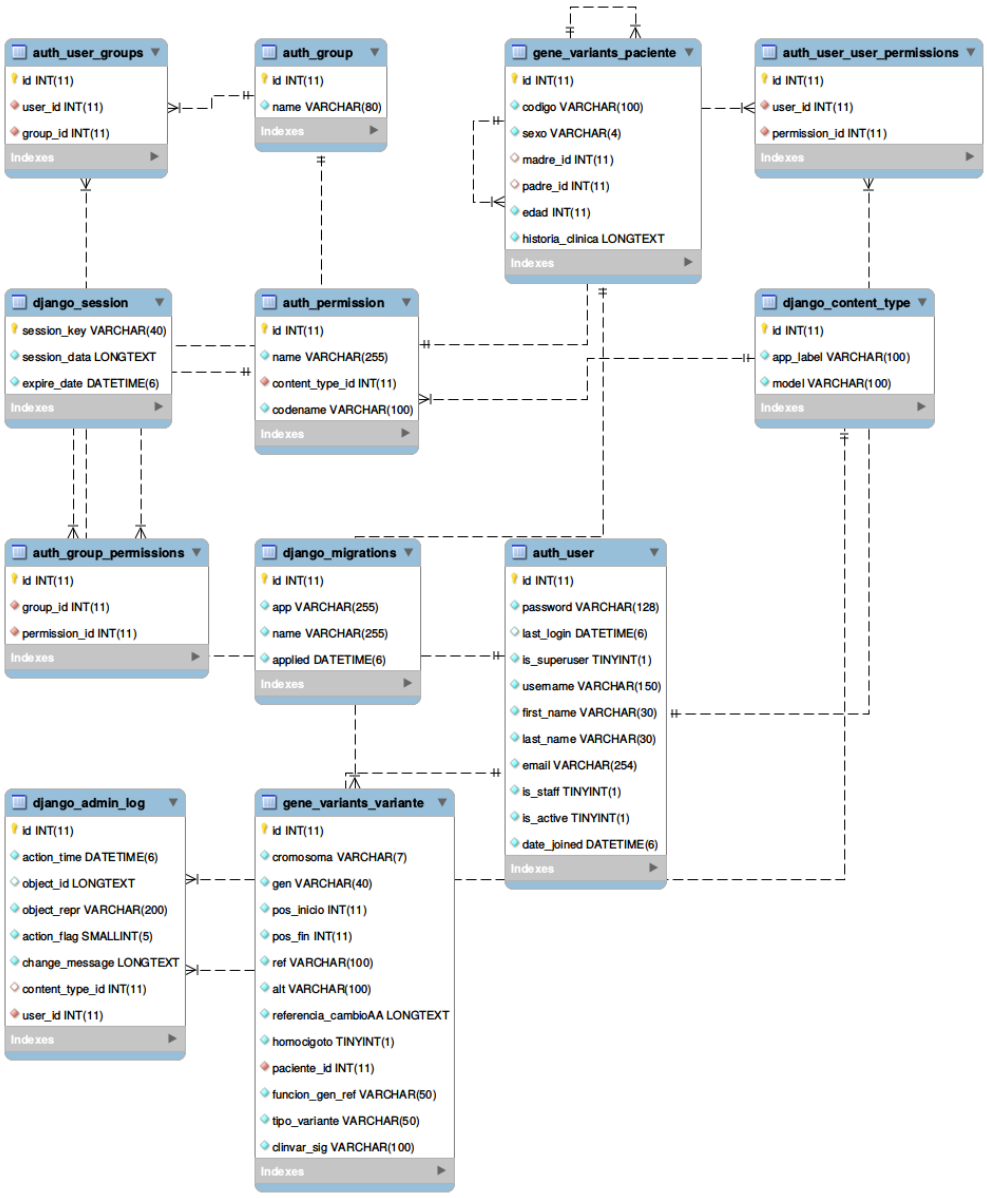


Figura 4.5.: Modelo entidad relación

La figura (4.5) muestra el modelo EER que se genero en la base de datos MySQL, donde se muestra el modelo EER que muestra las tablas generadas por la aplicación para crear

la base de datos donde se incluyen las tablas de para la gestión del las variantes junto con la historia clínica y se tiene encuentra la posible relación parental, padre, madre e hijo, el control de las consultas de la variantes indexadas y dentro de la misma base la gestión de accesos a otros usuarios con sus permisos.

4.4. Discusión

4.4.1. Gestión de datos biológicos

La importancia de gestión aplicada al manejo de datos clínicos y de información genética es de vital importancia dado que existen miles de anotaciones que requieren de scripts para cargarlos las anotaciones y como es este caso el historial clínico del paciente [26].

La aplicación desarrollada para crear y gestionar una base de datos aplicada una bioinformática con aplicaciones a la medicina, es necesario que la base de datos provea las consultas para soportar las decisiones sobre un paciente en especifico teniendo en cuenta sus datos,la relación con datos de otros pacientes y los datos de exomas, además de los datos relacionados con los familiares en caso de que se encuentren estos datos. Mostrando que es posible realizar una integración adecuada de los datos bioinformáticos y clínicos utilizando bases de datos relacionales, con una buena respuesta en las consultas. [27].

4.5. Conclusión

La utilización de aplicaciones en Django permite que un bioinformático diseñe e implemente bases de datos aplicadas al diagnostico clínico, donde se puede guardar y gestionar toda la información obtenida de un paciente, lo que permite hacer análisis a profesionales Médicos y biólogos fácilmente. Una vez ha sido implementa la base de datos también es posible aplicar técnicas de minería de datos para optimizar los análisis de los pacientes.

5. Capítulo 3

Se deben incluir tantos capítulos como se requieran; sin embargo, se recomienda que la tesis o trabajo de investigación tenga un mínimo 3 capítulos y máximo de 6 capítulos (incluyendo las conclusiones).

6. Capítulo ...

Se deben incluir tantos capítulos como se requieran; sin embargo, se recomienda que la tesis o trabajo de investigación tenga un mínimo 3 capítulos y máximo de 6 capítulos (incluyendo las conclusiones).

7. Conclusiones y recomendaciones

7.1. Conclusiones

Las conclusiones constituyen un capítulo independiente y presentan, en forma lógica, los resultados de la tesis o trabajo de investigación. Las conclusiones deben ser la respuesta a los objetivos o propósitos planteados. Se deben titular con la palabra conclusiones en el mismo formato de los títulos de los capítulos anteriores (Títulos primer nivel), precedida por el numeral correspondiente (según la presente plantilla).

7.2. Recomendaciones

Se presentan como una serie de aspectos que se podrían realizar en un futuro para emprender investigaciones similares o fortalecer la investigación realizada. Deben contemplar las perspectivas de la investigación, las cuales son sugerencias, proyecciones o alternativas que se presentan para modificar, cambiar o incidir sobre una situación específica o una problemática encontrada. Pueden presentarse como un texto con características argumentativas, resultado de una reflexión acerca de la tesis o trabajo de investigación.

A. Anexo: Nombrar el anexo A de acuerdo con su contenido

Los Anexos son documentos o elementos que complementan el cuerpo de la tesis o trabajo de investigación y que se relacionan, directa o indirectamente, con la investigación, tales como acetatos, cd, normas, etc.

B. Anexo: Nombrar el anexo B de acuerdo con su contenido

A final del documento es opcional incluir índices o glosarios. Éstos son listas detalladas y especializadas de los términos, nombres, autores, temas, etc., que aparecen en el mismo. Sirven para facilitar su localización en el texto. Los índices pueden ser alfabéticos, cronológicos, numéricos, analíticos, entre otros. Luego de cada palabra, término, etc., se pone coma y el número de la página donde aparece esta información.

C. Anexo: Nombrar el anexo C de acuerdo con su contenido

MANEJO DE LA BIBLIOGRAFÍA: la bibliografía es la relación de las fuentes documentales consultadas por el investigador para sustentar sus trabajos. Su inclusión es obligatoria en todo trabajo de investigación. Cada referencia bibliográfica se inicia contra el margen izquierdo.

La NTC 5613 establece los requisitos para la presentación de referencias bibliográficas citas y notas de pie de página. Sin embargo, se tiene la libertad de usar cualquier norma bibliográfica de acuerdo con lo acostumbrado por cada disciplina del conocimiento. En esta medida es necesario que la norma seleccionada se aplique con rigurosidad.

Es necesario tener en cuenta que la norma ISO 690:1987 (en España, UNE 50-104-94) es el marco internacional que da las pautas mínimas para las citas bibliográficas de documentos impresos y publicados. A continuación se lista algunas instituciones que brindan parámetros para el manejo de las referencias bibliográficas:

~~Institución~~	Disciplina de aplicación
Modern Language Association (MLA)	Literatura, artes y humanidades
American Psychological Association (APA)	Ambito de la salud (psicología, medicina) y en general en todas las ciencias sociales
Universidad de Chicago/Turabian	Periodismo, historia y humanidades.
AMA (Asociación Médica de los Estados Unidos)	Ambito de la salud (psicología, medicina)
Vancouver	Todas las disciplinas
Council of Science Editors (CSE)	En la actualidad abarca diversas ciencias
National Library of Medicine (NLM) (Biblioteca Nacional de Medicina)	En el ámbito médico y, por extensión, en ciencias.
Harvard System of Referencing Guide	Todas las disciplinas
JabRef y KBibTeX	Todas las disciplinas

Para incluir las referencias dentro del texto y realizar lista de la bibliografía en la respectiva sección, puede utilizar las herramientas que Latex suministra o, revisar el instructivo desa-

rrollado por el Sistema de Bibliotecas de la Universidad Nacional de Colombia[1], disponible en la sección "Servicios", opción "Trámitesz enlace .Entrega de tesis".

[1]Ver: www.sinab.unal.edu.co

Bibliografía

[1] Emad a. Mohammed, Behrouz H. Far, and Christopher Naugler. Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. *BioData Mining*, 7(1):1–23, 2014.

[2] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M. Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5(December):17875, 2015.

[3] Jiaxin Wu, Yanda Li, and Rui Jiang. Integrating Multiple Genomic Data to Predict Disease-Causing Nonsynonymous Single Nucleotide Variants in Exome Sequencing Studies. *PLoS Genetics*, 10(3), 2014.

[4] Madhuri Hegde, Avni Santani, Rong Mao, Andrea Ferreira-Gonzalez, Karen E. Weck, and Karl V. Voelkerding. Development and validation of clinical whole-exome and whole-genome sequencing for detection of germline variants in inherited disease. *Archives of Pathology and Laboratory Medicine*, 141(6):798–805, 2017.

[5] Yixue Li and Luonan Chen. Big Biological Data: Challenges and Opportunities Expanding volume of the big biological data and its bonanza. *Genomics, Proteomics & Bioinformatics*, 12(5):187–189, 2014.

[6] David Lauzon, Beatriz Kanzki, Victor Dupuy, Alain April, Michael S. Phillips, Johanne Tremblay, and Pavel Hamet. Addressing Provenance Issues in Big Data Genome Wide Association Studies (GWAS). *Proceedings - 2016 IEEE 1st International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2016*, pages 382–387, 2016.

[7] Coriell Institute. 1000 genomes project.

[8] Juan Felipe Arias-blanco, Dora Janeth Fonseca-mendoza, and Oscar Gamboa-garay. FRECUENCIA DE MUTACIÓN Y DE VARIANTES DE SECUENCIA PARA LOS GENES BRCA1 Y BRCA2 EN UNA MUESTRA DE MUJERES COLOMBIANAS CON SOSPECHA DE SÍNDROME DE CÁNCER DE MAMA HEREDITARIO: SERIE DE CASOS. *Revista Colombiana de Obstetricia y Ginecología*, 65(4):287–296, 2015.

[9] Quan Li and Kai Wang. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American Journal of Human Genetics*, 100(2):267–280, 2017.

[10] Kathleen M Fisch, Tobias Meißner, Louis Gioia, Jean Christophe Ducom, Tristan M Carland, Salvatore Loguercio, and Andrew I Su. Omics Pipe: A community-based framework for reproducible multi-omics data analysis. *Bioinformatics*, 31(11):1724–1728, 2015.

[11] Curtis Huttenhower and Oliver Hofmann. A quick guide to large-scale genomic data mining. *PLoS Computational Biology*, 6(5):1–6, 2010.

[12] Martine Tetreault, Eric Bareke, Javad Nadaf, Najmeh Alirezaie, and Jacek Majewski. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. *Expert Review of Molecular Diagnostics*, 2015.

[13] Geraldine A Van Der Auwera, Mauricio O Carneiro, Chris Hartl, Ryan Poplin, Ami Levy-moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V Garimella, David Altshuler, Stacey Gabriel, and Mark A Depristo. *From FastQ data to high confidence varant calls: the Genonme Analysis Toolkit best practices pipeline*, volume 11. 2014.

[14] Riyue Bao, Lei Huang, Jorge Andrade, Wei Tan, Warren a Kibbe, Hongmei Jiang, and Gang Feng. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. *Libertas Academica*, 13:67–82, 2014.

[15] Adam Cornish and Chittibabu Guda. A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International*, 2015(BioMed Research International):11, 2015.

[16] Babraham Bioinformatics. FASTQC manual, 2016.

[17] H Yang and K Wang. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc*, 10(10):1556–1566, 2015.

[18] Mehdi Pirooznia, Melissa Kramer, Jennifer Parla, Fernando S Goes, James B Potash, W Richard McCombie, and Peter P Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):14, 2014.

[19] Qian Zhou, Xiaoquan Su, Anhui Wang, Jian Xu, and Kang Ning. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE*, 8(4), 2013.

[20] Ram Vinay Pandey, Stephan Pabinger, Albert Kriegner, and Andreas Weinhäusel. ClinQC: a tool for quality control and cleaning of Sanger and NGS data in clinical research. *BMC Bioinformatics*, 17(1):56, 2016.

[21] Jason O'Rawe, Tao Jiang, Guangqing Sun, Yiyang Wu, Wei Wang, Jingchu Hu, Paul Bodily, Lifeng Tian, Hakon Hakonarson, W Evan Johnson, Zhi Wei, Kai Wang, and Gholson J Lyon. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine*, 5(3):28, 2013.

[22] Charles D. Warden, Aaron W. Adamson, Susan L. Neuhausen, and Xiwei Wu. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, 2:e600, 2014.

[23] Ellen Tsai, Rimma Shakbatyan, Jason Evans, Peter Rossetti, Chet Graham, Himanshu Sharma, Chiao-Feng Lin, and Matthew Lebo. Bioinformatics Workflow for Clinical Whole Genome Sequencing at Partners HealthCare Personalized Medicine. *Journal of Personalized Medicine*, 6(1):12, 2016.

[24] Xiangtao Liu, Shizhong Han, Zuoheng Wang, Joel Gelernter, and Bao Zhu Yang. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE*, 8(9):1–11, 2013.

[25] Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26, 2016.

[26] Umadevi Paila, Brad A. Chapman, Rory Kirchner, and Aaron R. Quinlan. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Computational Biology*, 9(7), 2013.

[27] W Sujansky. Heterogeneous database integration in biomedicine. *Journal of biomedical informatics*, 34(2001):285–298, 2001.