

Asociación de variantes en regiones codificantes de genes con datos clínicos en pacientes colombianos usando minería de datos.

Jennifer Vélez Segura

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

Asociación de variantes en regiones codificantes de genes con datos clínicos en pacientes colombianos usando 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
2018

(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 Bibliotecas¹, 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

¹ver: 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

Agradecimientos	VII
Resumen	IX
Lista de figuras	3
Lista de tablas	5
1 Introducción	6
2 Estado del Arte	10
2.1 Biología molecular y secuenciación masiva.	10
2.2 Bioinformática	13
2.2.1 Integración de datos genómicos y clínicos	14
2.2.2 Análisis de datos genómicos con aplicaciones clínicas	18
2.2.3 Minería de datos genómicos	20
3 Validación de un pipeline para la identificación de variantes	22
3.1 Identificación de variantes	22
3.2 Datos	23
3.3 Estrategias del pipeline	24
3.4 Resultados y validación	26
3.5 Discusión	33
3.6 Conclusiones	37
4 Modelo de integración de la información	38
4.1 Diseño e implementación del modelo de datos	39
4.2 Gestión de datos genómicos y clínicos	41
4.3 Conclusión	43
5 Modelo de minería de datos	44
5.1 Análisis exploratorio de los datos.	44
5.2 Análisis textual de información clínica.	46
5.2.1 Preprocesamiento.	46
5.2.2 Grupos de características clínicas.	47

Contenido	1
5.3 Asociación de grupos con variantes.	51
5.3.1 Variantes vistas como transacciones.	52
5.3.2 Experimentación	52
5.3.3 Resultados	53
6 Capítulo ...	64
7 Conclusiones y recomendaciones	65
7.1 Conclusiones	65
7.2 Recomendaciones	65
Bibliografía	66

Lista de Figuras

2.1.	SNPs en Humanos.	12
3.2.	Pipeline basado en las buenas practicas para el llamado de variantes.	24
3.1.	Pipeline basado en las buenas practicas para el llamado de variantes.	25
3.3.	Calidad del llamado de bases en una secuencia Estadísticas básicas del reporte FASTQ	26
3.4.	Variantes obtenidas por Omics pipe	27
3.5.	Variaciones de la muestra	28
3.6.	Distribución de variantes a lo largo de los cromosomas	28
3.7.	Distribución de variantes a lo largo de los cromosomas	29
3.8.	Diagrama de relación entre variantes comunes de Omics y de Illumina	29
3.9.	Distribución de variantes a lo largo de los cromosomas para los exomas.	30
3.10.	Variaciones de la muestras dentro del exoma	31
3.11.	Distribución de variantes a lo largo de los cromosomas para los exomas	31
3.12.	Diagrama de relación entre las variantes publicas y las obtenidas por el pipeline.	32
3.13.	Imagen de la variante presente en el exoma público	33
4.1.	Esquema de datos integrados	39
4.2.	Modelo entidad relación	40
4.3.	Interfaz de ingreso para administrar la base de datos.	41
4.4.	Interfaz de administración.	41
4.5.	Ingreso de pacientes.	42
4.6.	Consulta a variantes	42
5.1.	Distribución de rango de edades y géneros de los pacientes	45
5.2.	Distribución del tipo de variantes	45
5.3.	Frecuencias sin stop words y palabras sinónimas	47
5.4.	TF-IDF	48
5.5.	Número optimo de clusters	49
5.6.	Valor Silhouette por cada cluster	50
5.7.	Cluster 1	53
5.8.	Reglas de asociación del cluster 1 con variantes sinónimas.	54
5.9.	Reglas de asociación del cluster 1 sin variantes sinónimas.	55
5.10.	Cluster 2	55

5.11. Reglas de asociación del cluster 2 con variantes sinónimas.	56
5.12. Reglas de asociación del cluster 2 con variantes sinónimas.	56
5.13. Cluster 3	57
5.14. Reglas de asociación del cluster 3 con variantes sinónimas.	57
5.15. Reglas de asociación del cluster 3 sin variantes sinónimas.	58
5.16. Cluster 4	58
5.17. Reglas de asociación del cluster 4 con variantes sinónimas.	59
5.18. Reglas de asociación del cluster 4 sin variantes sinónimas.	59
5.19. Cluster 5	60
5.20. Reglas de asociación del cluster 5 con variantes sinónimas	60
5.21. Reglas de asociación del cluster 5 sin variantes sinónimas	61
5.22. Reglas de asociación con variantes sinónimas	62
5.23. Reglas de asociación sin variantes sinónimas	63

Índice de cuadros

2-1. SOFTWARE PARA LA INTEGRACIÓN DE DATOS GENÓMICOS CON FINES DE INVESTIGACIÓN	15
2-2. SOFTWARE DE INTEGRACIÓN DE DATOS GENÓMICOS CON FINES DIAGNÓSTICOS	17
2-3. SOFTWARE DE INTEGRACIÓN DE VARIANTES CON ENFERMEDADES	18
3-1. Tabla de Variantes obtenidas.	27
3-2. Tabla de Variantes obtenidas a partir de un exoma.	30
3-3. Tabla de validación 1.	32
3-4. Tabla de validación 2.	32

1 Introducción

El desarrollo de este trabajo responde a la necesidad actual del país en cuanto a la utilización las nuevas 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 métodos de análisis de datos a gran escala permitiendo mostrar un acercamiento de la estructura genética de la población colombiana asociada a la información clínica disponible 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 pronóstico de enfermedades presentes en la población colombiana aprovechando todas las avances que se encuentran a disposición actualmente y que permiten generar nuevos aportes con un impacto real en la salud.

En los últimos años con el desarrollo de las tecnologías NGS (Secuenciación 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 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 específico [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 cantida-

des 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 biológicas es una gran limitante dado que no tienen fundamentos de programación ni conocen los procedimientos que se utilizan normalmente en las ciencias de la computación, por lo tanto prefieren utilizar herramientas más 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 gráficas que consumen más 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 específicas con respecto a la información genómica [11]. Un ejemplo de esto es la utilización de algoritmos de agrupamiento para encontrar grupos de genes que están fuertemente relacionados con estados de evolución de los diferentes estadios en cáncer [5].

La gran cantidad de datos biológicos que se encuentran disponibles en la actualidad, deben ser tenidos en cuenta para la investigación y uso en el diagnóstico clínico, sin embargo estos datos presentan los siguientes inconvenientes para su acceso y disponibilidad como son: el almacenamiento, el procesamiento, la conexión y el análisis integrado de los mismos [12]. Por esta razón, en los procesos de diagnóstico se hace necesario reconocer los patrones sintomáticos de pacientes de los pacientes y asociarlos con variantes genéticas, ya que no ha sido

posible realizarlo fácilmente por diversas causas como la perdida de información relevante el acceso a la misma, por ejemplo los datos que reposan en las historias clínicas, que para su acceso se requiere solicitar un permiso [12].

La importancia de entender las formas genéticas que pueden causar diversos síndromes y patología, pueden permitir que se le dé prioridad diagnostica y de tratamiento a los pacientes afectados, teniendo en cuenta que aún se continua descubriendo nuevos genes asociados a enfermedades, teniendo en cuenta que existen retos computacionales como la integración de datos heterogéneos y de grandes cantidades que lleva a la necesidad de aplicar metodologías “big data” y minería de datos para análisis de datos génomicos [13, 14, 15]. La diversidad de los datos permite que la utilización de técnicas de minería de datos se pueda aprovechar para dar respuesta al problema de asociación entre variantes genéticas y el impacto clínico de esas variantes [12].

Fuentes de información

Los datos clínicos y genómicos fueron donados por el **Centro de Investigación en Genética Humana y Reproductiva Genetix S.A.S** que es dirigido por la Dra. *Claudia Serrano Serrano M.D. MSc.*

Actividades desarrolladas

En el presente trabajo se desarrollaron las siguientes actividades:

1. La implementación y validación de un pipeline para la identificación de variantes.
2. El diseño e implementación de un sistema de información para realizar la gestión de datos para las variantes obtenidas junto con la información clínica.
3. Diseñar e implementar un modelo para la minería de datos aplicada en pacientes colombianos.
4. Visualizar y validar los resultados obtenidos a partir del modelo de minería de datos.

Objetivos

Objetivo General

Proponer un modelo de minería de datos para la asociación de variantes identificadas en regiones codificantes de genes con datos clínicos que apoyen el diagnóstico en pacientes.

Objetivos Específicos

1. Implementar una estrategia de pipeline para la identificación de variantes en regiones codificantes de 4813 genes de una muestra poblacional.
2. Diseñar e implementar modelo de datos que permita la integración de la información de las variantes en regiones codificantes y la información clínica disponible de una muestra poblacional.
3. Diseñar e implementar el modelo de minería de datos que permite la asociación entre las variantes identificadas en regiones codificantes de genes con datos clínicos en pacientes colombianos.
4. Validar y visualizar los resultados del modelo implementado con las variantes identificadas en regiones codificantes de 4813 genes en pacientes colombianos asociados a los datos clínicos.

Contribuciones

- * Validación de un pipeline para el llamado de variantes en exones humanos. Presentado en el Congreso de Genética Humana.Cali-Colombia 2016.
- * Implementation of a pipeline for the identification of variants and an information management system in Colombian patients. Escuela Latinoamericana de Genética Humana y Médica - ELAG. Caxias do Sul, RS, Brasil 2017.
- * Sistema de consulta de variantes en una muestra de pacientes colombianos.
- * Exomic and clinical data management using Django en IV Congreso Colombiano de Bioinformática y Biología Computacional y la VIII Conferencia Iberoamericana de Bioinformática. Cali-Colombia 2017.
- * Modelo de minería en datos genómicos y clínicos. Conferencia Pycon Colombia.Medellín 2018.

Estructura del documento

El presente trabajo se distribuye en los siguientes capítulos: 1.Estado del arte, 2.Pipeline para la identificación de variantes, 3.Modelo para integración de la información, 4. Modelo de minería de datos genómicos, 5. Conclusiones,recomendaciones y trabajo futuro, 6. Bibliografía y anexos.

2 Estado del Arte

Con el desarrollo de las tecnologías de secuenciación masiva los biólogos moleculares se han visto en la necesidad de utilizar metodologías computacionales para analizar datos biológicos a gran escala, además de la aplicación de estas tecnologías en medicina requieren de un diseño y una validación que permita obtener nuevos conocimientos que sean aplicables en la salud de los colombianos.

2.1. Biología molecular y secuenciación masiva.

Desde que Watson y Crick propusieron la estructura del ADN en 1953 [16], el estudio del ADN ha sido básico en el desarrollo de la biología molecular, incluso el mismo Francis Crick fue quien propuso el dogma central de la misma para describir la relevancia del ADN en los seres vivos y la utilización de la información que contiene por las células, dada la importancia del ADN en las décadas de 1970 y 1980 se desarrollaron técnicas para determinar el orden de los nucleótidos (técnicas de secuenciación) de manera más eficiente que la secuenciación de las proteínas y se definieron secuencias de algunos organismos como el virus de Epstein Barr y de la mitocondria humana, mediante la utilización de métodos químicos propuestos por Maxam y Gilbert en 1977 y Sanger en 1980 siendo este último el más popular, estas técnicas son conocidas como técnicas primera generación [17].

Dada la importancia del ADN en las décadas de 1970 y 1980 se desarrollaron técnicas para determinar el orden de los nucleótidos (técnicas de secuenciación) de una manera más eficiente que la secuenciación de las proteínas y se definieron secuencias de algunos organismos como el virus de Epstein Barr y de la mitocondria humana, mediante la utilización de métodos químicos propuestos por Maxam y Gilbert en 1977 y Sanger en 1980 siendo este último el más popular, estas técnicas son conocidas como técnicas primera generación [17].

Los métodos desarrollados para secuenciar prosperaron y con el proyecto del genoma humano que comenzó en 1980 y fue completado en el 2003, permitió que se desarrollaran nuevas tecnologías para optimizar el proceso de secuenciación y disminuir sus costos, inicialmente fue el secuenciador de Illumina que en el 2008 permitió obtener el primer individuo humano secuenciado con esta tecnología [18]. Estas nuevas tecnologías se fueron desarrollando en otras plataformas, tales como el secuenciador de roche 454 y el SOLiD de applied biosistens

[18] y son conocidas como tecnologías de última generación o de siguiente generación (Next-generation sequencing, NGS), que tienen la capacidad de realizar secuenciaciones de alto rendimiento de una manera más rápida y económica que las de primera generación [17]. La diferencia entre las técnicas de secuenciación de primera generación y las de NGS se presenta en el hecho de que la nueva generación genera lecturas de menos de 500 pares de bases en comparación a las 1000 pares de bases de sanger [18, 19].

El desarrollo de estas tecnologías ha hecho que los datos genómicos aumenten de una manera vertiginosa, y permiten que se pueda realizar análisis en diferentes organismos con aplicaciones en biotecnología y salud [17]. Se ha estimado que cada genoma humano tiene alrededor de 3.5 millones de diferencias con respecto al genoma de referencia (Genoma de consenso para la salud humana), estas diferencias son llamadas variantes y pueden determinar el fenotipo de los individuos, algunas de estas variantes son conocidas para indicar predisposiciones a enfermedades [20].

En el campo de la salud actualmente se emplea la secuenciación de exomas es la más utilizada puesto que se considera que los exones son las regiones de ADN conservadas y expresadas, (se traducirán en ARNm y posteriormente en proteínas) y representan menos del 2% el genoma humano pero se estima que contiene el 85% de las variantes conocidas de enfermedades, lo que permite la reducción de costos y una buena alternativa frente a la secuenciación de genomas completos [21, 22, 17].

Dado que la utilización de NGS permitió dar respuesta para entender enfermedades raras, ya que fue posible identificar las regiones responsables de una enfermedad, teniendo como control los datos poblacionales como el proyecto de 1000 genomas, los datos genómicos pueden también ser útiles para la caracterización de enfermedades poligénicas y su asociación con las variaciones genómicas presentes en el individuo [23].

La secuenciación de exones (secuenciación de exoma) ha sido un buen método para identificar SNPs (Polimorfismos de Nucleótido Único) y los SNV (Variantes de nucleótido simple) como se observa en la siguiente figura 2.1, y permitió identificar pequeñas delecciones o inserciones (indels) que pueden ser la causa de enfermedades y de la variación en los fenotipos [24, 25].

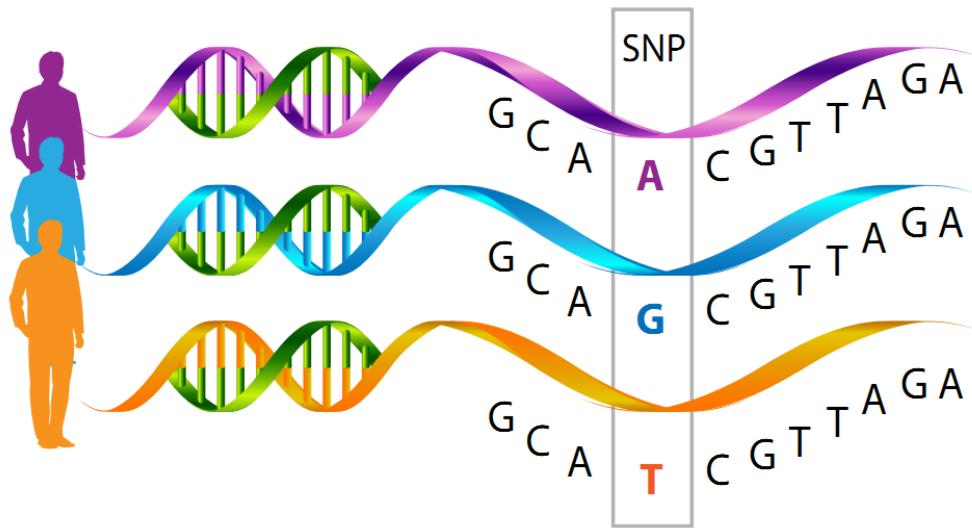


Figura 2.1: SNPs en Humanos.

A partir de la información pública disponible se ha estimado que las variantes pueden afectar la función de la proteína ya su vez pueden estar asociados a otros genes dentro de una enfermedad, aunque en genética humana la frecuencia de una variante específica dentro de la población es clasificada como benigna (No genera una enfermedad), por ello la importancia de la integración de la información y la revisión de la variante [26].

La identificación de variantes se realiza con diversas plataformas siendo el MiSeq es un sistema de secuenciación de illumina que permite la secuenciación de 4800 genes en un solo experimento y una de las más populares [21]. En cuanto al análisis de los datos esta plataforma incluye un servicio de computación en la nube (BaseSpace), donde los datos biológicos son analizados sin necesidad de que los investigadores tengan habilidades en bioinformática. Pero están disponibles como herramientas solamente de investigación y no como diagnóstico [21].

A pesar de ser herramientas de investigación que han sido desarrolladas, Illumina permite que dentro del BaseSpace se publiquen nuevos algoritmos, herramientas abiertas o aplicaciones diseñadas por desarrolladores que permitan mejorar estos análisis genómicos y la aplicación en diversas en sus diversas plataformas de secuenciación que ofrece esta empresa [21].

Los datos obtenidos a partir de técnicas de NGS han tenido un crecimiento vertiginoso y presentan un reto para el manejo y análisis de los mismos, debido a que los formatos de los datos y las inconsistencias de las secuencias como resultado de los procesos experimentales, la importación de las secuencias a nivel digital, el ensamblaje de los fragmentos de ADN, el alineamiento y post-alineamiento de grandes cantidades de datos biológicos hace que se con-

vienta en una de las bases de la investigación en bioinformática [24, 27].

La secuenciación de siguiente generación (NGS) aplicaciones en salud

La secuenciación de siguiente generación ha sido adoptada en el ámbito clínico, dado que se ha documentado su utilidad para el diagnóstico de enfermedades y para la toma de decisiones en cáncer o para la selección de dosis de medicamentos en un paciente, [28] algunos de estos ejemplos son:

Cáncer de Seno

El cáncer de seno es una enfermedad que afecta principalmente a mujeres y que en estadios avanzados tienen una alta tasa de mortalidad por lo que ha recibido una importante atención de por la comunidad de investigadores, principalmente en el área biomédica con la intención de buscar marcadores genéticos de la enfermedad. Actualmente se encuentra una gran cantidad de investigaciones publicadas, donde se intenta visualizar las interacciones de los genes como esos marcadores en las distintas poblaciones [29].

Teniendo en cuenta que el cáncer es el resultado de una mutación de ADN, donde la consecuencia es que la célula portadora de la mutación pierda su función normal y gane la habilidad de multiplicarse de manera indefinida sobre los tejidos normales. Donde la identificación más común para realizar la identificación de biomarcadores genéticos es la utilización de NGS, donde la variación de un gen puede alterar la función celular y ser causal de la enfermedad y en algunos casos puede ser heredable y predisponente al desarrollo de la enfermedad [29, 25].

Fibrosis Quística

La fibrosis quística es una enfermedad multisistémica causada por mutaciones puntuales en el gen CFTR, las características típicas de esta enfermedad son: la enfermedad pulmonar obstructiva, infecciones bacterianas crónicas de las vías respiratorias y senos paranasales e infertilidad masculina debida a azoospermia obstructiva, la mayoría de los pacientes con esta enfermedad tienen insuficiencia pancreática, es frecuente que los pacientes con fibrosis quística tengan mutaciones en el gen CFTR con un efecto funcional de la proteína leve, se han identificado 2000 variantes asociados a esta enfermedad [30].

2.2. Bioinformática

La bioinformática según la asociación americana de patología y el colegio americano de patología es la disciplina que conceptualiza la biología en términos de macro-moléculas y aplica técnicas informáticas (matemática aplicada, ciencias de la computación y estadística)

para entender y organizar la información asociada a esas macro-moléculas, en gran escala [31].

La bioinformática combina retos de investigación en las áreas de la biología y la informática para desarrollar diferentes métodos y herramientas para el análisis de datos biológicos y puede tratar acerca del almacenamiento, simulación y análisis de datos biológicos aplicando el uso de herramientas computacionales como la minería de datos, esta ultima siendo definida como una herramienta de investigación, desarrollo y aplicación para expandir el uso de los datos biológicos y médicos con fines de investigación y generación de nuevos conocimientos, incluyendo las herramientas que permitan almacenar, archivar y analizar o visualizar dichos datos [32]

El auge de las tecnología de NGS permitió que la bioinformática diera respuesta a las dificultades que presenta la genómica en la búsqueda de ser una nueva innovación biomédica y en otras áreas de las ciencias biológicas, donde el valor de la bioinformática radica en la promesa de que la información genómica tiene grandes beneficios que son aplicables al área de la salud aunque estos la obtención de información relevante presentan un varios retos uno de ellos es la integración de los datos genómicos y clínicos y los derechos de propiedad sobre los mismos[33].

2.2.1. Integración de datos genómicos y clínicos

En la era de las omáticas, los datos se presentan en diferentes formas y en varios niveles en términos biológicos, los cuales incluyen los datos genómicos, datos de transcriptomica, epigenomica, metabolomica, entre otros, donde se incluyen también las diferentes datos poblacionales humanos y las historias clínicas, la escala de estos datos se encuentran entre pentabyte y exabyte [5].Aunque la definición de “big data” es muy discutida dentro de las ciencias de la información, sin embargo el nombre se hace referencia a la “gran cantidad de datos” que se caracterizan por el volumen del procesamiento, la variabilidad de los mismos y la veracidad de la calidad de los datos [5]. Partiendo de lo anterior los datos genómicos pueden ser catalogados como “big data” ya que poseen las siguientes características: Son numerosos, no pueden ser almacenados dentro de una base regular de datos, la velocidad de generación y procesamiento es muy rápida [34].

En el diagnóstico de enfermedades los datos genómicos vistos como ”big data” comparten los mismos retos tecnológicos como son: el almacenamiento, la transferencia de la información, control del acceso y manejo de la información, otros retos computacionales propios de los datos es el moldeamiento de los sistemas biológicos, la gran escala y diversidad de los datos donde los modelos no optimizados que pueden fallar [35].

Para el manejo de los datos se han aplicado varios modelos de sistemas de información en

bioinformática con diversas herramientas para integrar datos biológicos, utilizando sistemas de bodega de datos que están disponibles de manera gratuita y que fueron desarrollados con el fin de dar respuesta algunos de los problemas en el manejo de datos biológicos, dada la importancia que tiene de poder utilizar toda la información necesaria de manera eficiente [27], algunas de estas bases de datos públicas son las de NCBI y ensambl que hacen parte de un consorcio internacional [36, 37]. En la tabla **2-1** se describen algunos softwares libres para la integración de datos.

Cuadro 2-1: SOFTWARE PARA LA INTEGRACIÓN DE DATOS GENÓMICOS CON FINES DE INVESTIGACIÓN

Software	Descripción
BioMart	Permite la integración de datos biológicos, esta herramienta optimiza de manera rápida la integración de grandes cantidades de datos, de fácil uso (los biólogos moleculares y médicos no poseen generalmente bases sólidas de programación) y ha sido usado por laboratorios para integrar portales de enfermedades de cáncer, datos de microarreglos y expresión génica [27].
BioXRT	Fue desarrollado por biólogos para publicar sus datos en internet, está recomendado para laboratorios pequeños y que necesiten publicar sus resultados para correlacionarlos con otros resultados de otros laboratorios, también ha sido utilizado en varios proyectos para la anotación del cromosoma 7 y el estudio estructural de variantes genéticas asociadas a autismo [27].
InterMine	Provee un modelo de datos para llamar 28 bases de datos libres como Gene Ontology, que permite la implementación de flujos de trabajos de manera automatizada [27].
PathwayTools	Es una herramienta que permite utilizar organismos específicos para realizar las búsquedas o modelos de organismos en las bases de datos, estas pueden ser publicadas y visualizadas en la web, incluye una predicción de varias reacciones metabólicas de los microorganismos [27].
Illumina	Presenta su propia herramienta integradora que permite hacer la anotación funcional de genes, la filtración y categorización de datos que puedan tener un impacto biológico en variantes relevantes, generando un reporte resumido de las enfermedades con un significado biológico y en un formato estructurado, pero no integra la información clínica de un paciente [21].

Muchas herramientas han sido implementadas con fines de investigación, más no con fines diagnósticos, en este sentido se han implementado otras herramientas que permiten integrar

datos con fines diagnósticos, ya que en este caso se requieren parámetros de seguridad por los datos que contienen información clínica y que deben ser manejados de manera privada. Esto implica otro manejo de datos biológicos ya que se adicionan nuevos datos como condiciones del paciente, tratamientos entre otros datos [38]. La tabla **2-3** presenta algunos de los softwares para integrar datos con fines de diagnósticos.

Cuadro 2-2: SOFTWARE DE INTEGRACIÓN DE DATOS GENÓMICOS CON FINES DIAGNÓSTICOS

Software	Descripción
BRISK: Biology-Related Information Storage kit	Es un paquete de recursos abiertos, permite relacionar una descripción fenotípica y una mutación somática (SNP), lo que permite a los investigadores proveer una asociación de estudios genómicos y capacidades de análisis, teniendo en cuenta el manejo de la muestra [27].
CaTRip	Fue desarrollada como un componente de caBIG, este software permite encontrar pacientes con perfiles similares, teniendo en cuenta el registro que hay dentro del sistema de datos clínicos, permite almacenar, cualificar y analizar datos de diferentes tipos de cáncer [38].
CBio Cancer Genomics Portal	Es otra herramienta que permite integrar datos definidos en la historia clínica de un paciente, como su descripción fenotípica, con la mayor cantidad de datos de ADN, ARNm, proteínas y de las imágenes obtenidas dentro de los diferentes exámenes realizados al paciente [38].
G-DOC Georgetown Database of Cancer	Fue desarrollada para integrar datos de las características de los pacientes con los datos biológicos, esta herramienta se enfoca en la visualización y análisis de datos [38].
iCOD Integrated Clinical Omics Database	Esta herramienta combina la patología clínica de los pacientes y la información molecular de pacientes con el fin de dar una información holística de los pacientes, fue desarrollado de manera local y permite la visualización de mapas de enfermedades que permite la interrelación clínica con los datos biológicos [38].
iDASH Integrating data for analysis, anonymization and sharing	No es una herramienta, pero si es un a infraestructura poderosa que permite la integración de datos y su análisis, distribuye herramientas y algoritmos enfocados en la privacidad de los datos [38].
tranSMART	Es una herramienta abierta que permite a los investigadores hacer relaciones entre el fenotipo y los datos moleculares, Da a los investigadores herramientas para generar descripciones y análisis estadísticos [38].

Otras herramientas han sido desarrolladas para encontrar asociaciones de variantes y genes afectados con las enfermedades requieren que se combinen los análisis de variantes con los

individuos donde se tenga acceso a la información de manera eficiente [20]. Algunas implementaciones desarrolladas para hacer esta tarea son:

Cuadro 2-3: SOFTWARE DE INTEGRACIÓN DE VARIANTES CON ENFERMEDADES

Software	Descripción
Variant-DataBase (Variant-DB) within	Es una base de datos implementada en PostgreSQL junto con Django para almacenar y manejar datos genómicos que se conectan a través de transSMART para asociar las variantes a un fenotipo [20].
HGVD	Es una herramienta con acceso web que permite manejar las variantes dentro de la población japonesa obtenidas a partir de secuenciación de exomas y genomas implementada en PostgreSQL y la interfaz gráfica con JBrowse [39].
Variome Project	Es un proyecto no gubernamental internacional que trabaja para integrar las variaciones genéticas y su efecto en la salud humana y que a su vez esta información sea curada, interpretada y compartida de manera gratuita [40].

2.2.2. Análisis de datos genómicos con aplicaciones clínicas

A nivel mundial se han clasificado los datos genómicos en cinco tipos que son de gran tamaño y que son ampliamente usados en la investigación en bioinformática, estos datos son: 1) Los de expresión génica, 2) datos de secuenciación de ADN, ARN y proteínas, 3) los de interacciones entre proteínas (PPI), 4) los de ruta metabólicas y 5) los datos de gene ontology (GO). Además se encuentran los datos de redes donde se asocian los genes con enfermedades que tienen una alta importancia en la investigación y el diagnóstico [41].

Dentro del análisis de datos de secuenciación los desarrollos se han enfocado en el manejo de la gran cantidad de información generada, mientras que en las asociaciones con enfermedad se enfocan en la asociación multi-objetivo entre la enfermedad y las redes heterogéneas son utilizados para establecer la relaciones entre los genes y la enfermedad; la complejidad de estas relaciones implican la utilización de herramientas de aprendizaje de máquina para reorganizar y visualizar la gran cantidad de datos obtenidos, y así poder realizar análisis y diagnóstico de enfermedades [41].

Dentro de las secuencias para el análisis a gran escala se ha utilizado la plataforma de Hadoop MapReduce, utilizando también BioPig como herramienta que se basa en el análisis de secuencias a nivel masivo utilizando la arquitectura de MapReduce, otra herramienta

está el Crossbow que se combina con Bowtie para dar una respuesta ultrarrápida con un uso eficiente de memoria para el alineamiento de lecturas cortas y SoapSNP que permite la identificación de SNP en genomas completos a través de computación en la nube o de manera local utilizando un clúster de hadoop. Otras herramientas basadas en la nube son Stormbow, CloVR y Rainbow. Otras plataformas que no utilizan herramientas de big data son Vmatch y SeqMonk [41].

Una de las herramientas más populares para el manejo el análisis de secuenciación de alto son Galaxy Project que permite el análisis de los diferentes tipos de datos por medio de una interfaz web o de manera local y es un software libre, también permite crear flujos de trabajo automatizados [17]. Otra herramienta es GATK que fue desarrollada por el Broad Institute y que se enfoca en el descubrimiento de variantes a diferentes niveles y con diversos organismos y con usos investigativos [18]. GATK a diferencia de Galaxy Project no tiene una interfaz gráfica y debe ser instalado en equipos con Linux y basa su arquitectura utilizando hadoop MapReduce para el procesamiento de los datos [13] .

Igualmente se han realizado implementaciones para análisis en bioinformática implementando los algoritmos de alineamiento múltiple en Hadoop y utilizando HBase, paralelizando la versión del NCBI del algoritmo BLAST, también se ha aplicado a nivel clínico la cantidad de datos producidos por los laboratorios como los record médicos electrónicos, datos biomédicos, datos biométricos, expresión génica entre otros y se ha utilizado el framework de MapReduce para realizar análisis simultáneo con un retorno rápido de resultados, haciendo que la promesa de que los análisis de “big data” en bioinformática y la salud sea aplicable [1].

Cada una de las herramientas han sido desarrolladas para responder al manejo datos en bioinformática y su análisis, Colombia se ha propuesto el usos de las bodegas de datos para dar soporte a la investigación, ya que el uso de estas metodologías han sido ampliamente aplicados en inteligencia de negocios, y se presenta la modelación multidimensional de datos biomédicos basados en bodega de datos [42].

Bustos [42] y colaboradores proponen que la bodega de datos aplicable en bioinformática es un híbrido entre Data Warehouse (bodega de datos) y data marts, utilizando la aplicación de descubrimiento de conocimiento (KDD) en los datos almacenados. El modelo propuesto es: 1) La selección de datos. 2) El agrupamiento y 3) Clasificación. En bioinformática se han aplicado las técnicas de minería para tratar de resolver diversos problemas biológicos, dependiendo del tipo de problema que se quiera abordar. Por ejemplo para la exploración de variantes de nucleótido simple (SNPs) asociados a enfermedades se ha implementado el algoritmo Apriori para buscar dentro de un set de atributos reglas que sean consistentes con la literatura, teniendo en cuenta que existen millones de SNPs que están correlacionados con varios fenotipos [43].

2.2.3. Minería de datos genómicos

La minería de datos biológicos (visto desde la bioinformática) es el proceso de extraer nuevo conocimiento (previamente desconocido) de datos biológicos, esto permite también la utilización de conceptos de minería de datos y aprendizaje de máquina en teorías y aplicaciones en la investigación biológica, dependiendo de los datos que se estén utilizando para ser aplicados, se encuentran los genómicos que provienen del secuenciación de ADN, los transcriptómicos que son de secuenciación de RNA o los de proteínas que provienen de las inferencias y los datos experimentales desde la química [44].

Las inferencias con respecto a las grandes cantidades de datos genómicos requieren análisis computacionales para interpretar los datos, siendo una de las áreas más activas de análisis donde se utiliza la minería de datos (entiendo la minería de datos como el método de extraer información por medio del aprendizaje de máquina, la estadística, la inteligencia artificial, patrones de reconocimiento y visualización) para resolver problemas biológicos, algunos ejemplos donde se ha aplicado técnicas de minería es la clasificación de genes, análisis de mutaciones en cáncer y expresión de genes [32].

También han sido aplicadas técnicas de agrupación de genes expresados diferencialmente, las máquinas de soporte vectorial han sido utilizados para asociar interacciones entre genes y generar redes biológicas, igualmente las metodologías tradicionales de minería de datos en ocasiones no son precisas o eficientes y requieren que se desarrollen nuevos algoritmos y metodologías que respondan de una manera más acertada a una pregunta biológica [45]. Sin olvidar que se requiere evaluar las plataformas disponibles, las herramientas tecnológicas que permitan la implementación de procesos que asocien los datos a la investigación y obtener resultados más generalizados. Esto debe estar aplicado a los requerimientos de los investigadores para garantizar una implementación exitosa [42, 45].

Algunas de las tareas de minería de datos son: 1. Clasificación: Donde se clasifican los datos a una clase predefinida, 2. Asociación: Ver elementos que están asociados mediante reglas, 3. El clustering o agrupamiento: Como la definición de una población de datos dentro de un subgrupo o cluster [32].

La utilización de las técnicas de secuenciación de alto rendimiento junto la aplicación de técnicas de minería de datos pueden aportar al diagnóstico de enfermedades complejas como las fallas cardíacas y el cáncer que presentan diversas causas [14]. Partiendo de lo anterior se hace necesario saber la relación entre las moléculas biológicas y las características de una enfermedad vistas desde la alteración de uno o varios genes y las posibles alteraciones que estos causan en una persona [5].

Resumen

Se presenta el estado del arte de la secuenciación de siguiente generación y el impacto que ha tenido en el diagnóstico, pronóstico y seguimiento de enfermedades complejas y las posibilidades de análisis y aplicaciones que puede traer el uso de esta tecnología, además de la necesidad de utilizar metodologías para analizar y obtener información relevante a partir de los datos genómicos.

3 Validación de un pipeline para la identificación de variantes

Los pipelines son un componente integral de la secuenciación de siguiente generación (NGS), para el procesamiento de los datos en bruto y detectar las alteraciones genómicas que tienen un impacto en la salud de un paciente, por lo tanto se hace necesario desarrollo, validación y monitoreo de los pipelines son necesarios para disminuir los errores de la identificación de variantes ya que pueden tener una consecuencia negativa en la salud de los pacientes [31]. En este capítulo se presenta el proceso para validar un pipeline que permitió la identificación de variantes a partir de datos de secuenciación de siguientes generación (NGS), donde se utiliza datos públicos un paciente para validar la calidad de las variantes, y se encuentra organizado de la siguiente forma 1. Identificación de variantes, 2. Datos. 3. Estrategias del pipeline. 4. Resultados y validación. 5. Discusión, 6.conclusiones 7. Resumen.

3.1. Identificación de variantes

Los pipelines bioinformaticos para NGS son comúnmente desarrollados en una plataforma específica y pueden ser adaptados según las necesidades del laboratorio, la mayoría de los pipelines consisten en los siguientes pasos [31]:

1. Generación de secuencias.
2. Alineamiento de las secuencias.
3. Llamado de variantes.
4. Filtrado de variantes.
5. Anotación de variantes.
6. Priorización de variantes.

La eficiencia de la identificación 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 % [46]. 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) [47].

Para el presente trabajo se realizaron mediciones de la calidad de las secuencias y el mapeo, post-alineamiento y el llamado de variantes, siguiendo las buenas prácticas para el llamado de variantes [10]. Teniendo en cuenta que existen múltiples herramientas para realizar el llamado de variantes tanto de uso privado como open source que permite seguir las buenas prácticas de identificación de variantes se hace necesario integrar las diversas herramientas para poder obtener los datos de buena calidad. 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 [48][49].

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

Para poder llevar a cabo 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].

3.2. Datos

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.

Para la validación del pipeline se corrió 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 research*). Se descargó el archivo bed para filtrar las variantes que se encuentran dentro del genoma completo de la muestra para obtener solo exones del NCBI para el genoma hg19. También se realizó una obtención de variantes a partir de un exoma completo público de la muestra NA12878, los datos fueron obtenidos vía ftp en la siguiente dirección:

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/

3.3. Estrategias del pipeline

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 [48].

Se utilizo como base modulo de omics-pipe propuesto por Fish y colaboradores [10] que presenta el pipeline que es acorde con las buenas practicas para el llamado de variantes en la figura 3.1

Para el presente trabajo se adiciono el filtrado específico de variantes según GATK y la parte de anotación de variantes con wAnnotar de la siguiente manera como muestra la siguiente figura 3.2:



Figura 3.2: Pipeline basado en las buenas prácticas para el llamado de variantes.

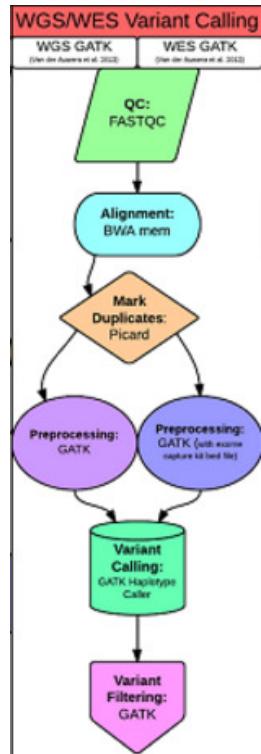


Figura 3.1: Pipeline basado en las buenas prácticas para el llamado de variantes.

Herramientas computacionales

Las herramientas bioinformáticas seleccionadas se implementaron en un clúster¹ que cuenta con las siguientes características:

- ⇒ Un nodo mestre 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 procesadores Intel Xeon E5-2695 – 24 cores 48 con HT / 192 GB RAM (4.378Tflops), 300 GB 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 Procesadores Intel Xeon X5690 - 12 cores / 192 GB RAM (3.659Tflops), 1.6 TB de Disco duro.

Y se instaló el módulo 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.

¹El clúster utilizado fue prestado por la Universidad de los Andes

3.4. Resultados y validación

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 [50].

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 3.3 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 sobre representadas.

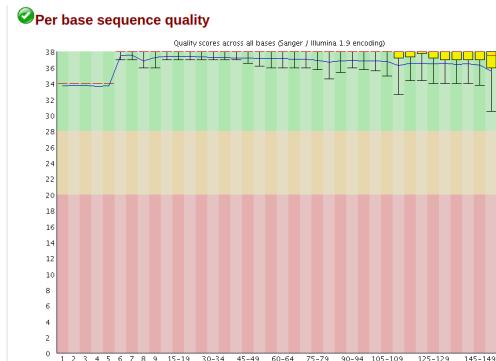


Figura 3.3: Calidad del llamado de bases en una secuencia Estadísticas básicas del reporte FASTQ

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_{PHRED}) que implementa el secuenciador [50]).

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

[47]. Dentro del pipeline solo se encuentra el proceso de llamado de variantes y no el proceso de filtrado de las mismas y que debió 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

Cuadro 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 3.4:

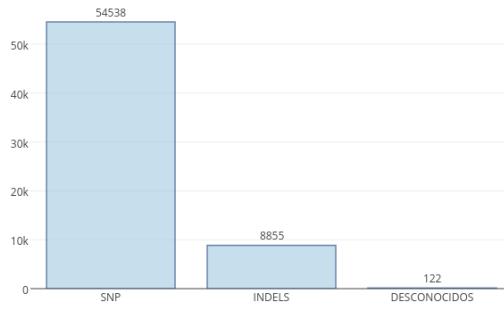


Figura 3.4: Variantes obtenidas por Omics pipe

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 3.5:

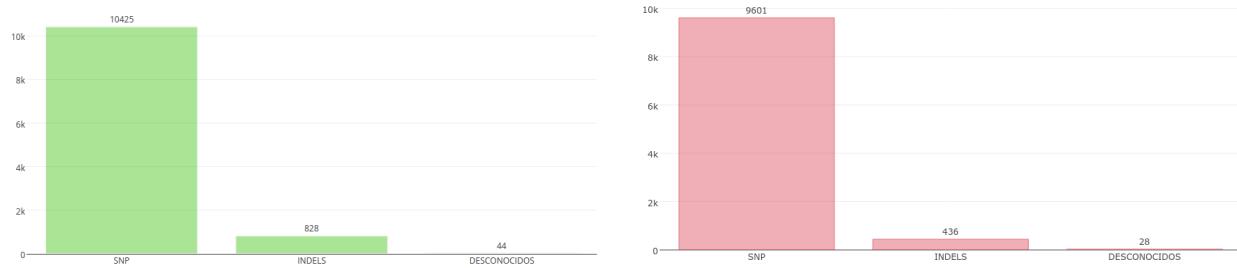


Figura 3.5: Variaciones de la muestra

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 3.6 y 3.7 :

Cromosoma	Variantes Omics	Variantes Calibradas	Variantes Illumina
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 3.6: Distribución de variantes a lo largo de los cromosomas

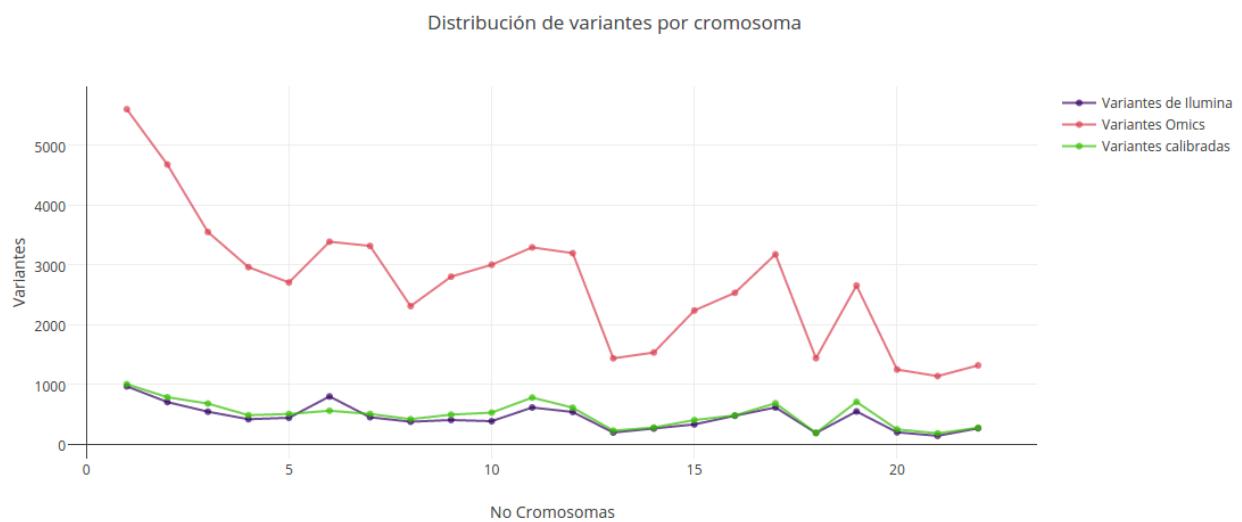


Figura 3.7: Distribución de variantes a lo largo de los cromosomas

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 cuando 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 (3.8):

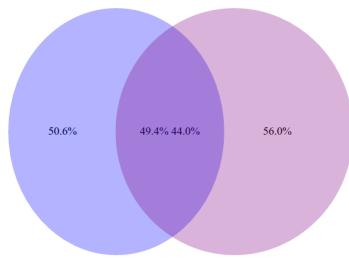


Figura 3.8: Diagrama de relación entre variantes comunes de Omics y de Illumina

Variantes de exoma vs variantes de omics

Una vez obtenidas las regiones se realizó 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

Cuadro 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 3.10

La distribución de las variantes a lo largo de los cromosomas se presenta en la siguiente tabla 3.9:

Cromosoma	Variantes públicas	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 3.9: Distribución de variantes a lo largo de los cromosomas para los exomas.

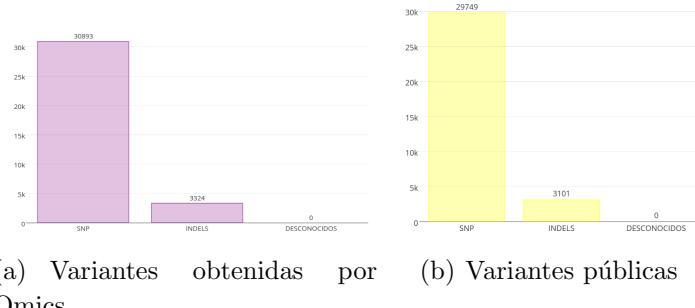


Figura 3.10: Variaciones de la muestras dentro del exoma

Y la representación gráfica de las variantes sobre la distribución a lo largo de los cromosomas figura 3.11:

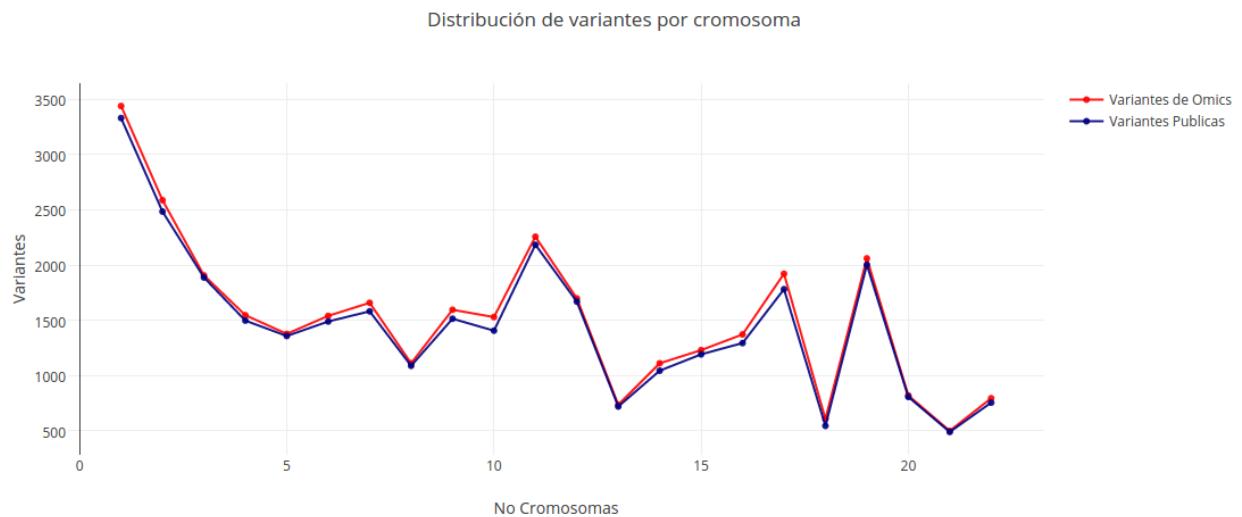


Figura 3.11: Distribución de variantes a lo largo de los cromosomas para los exomas

En la figura 3.11 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.

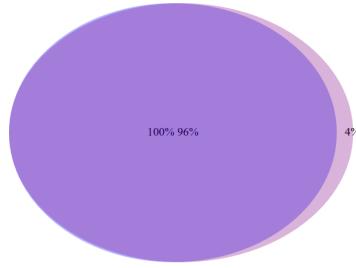


Figura 3.12: Diagrama de relación entre las variantes publicas y las obtenidas por el pipeline.

El diagrama de la figura 3.12 muestra la comparación de las variantes obtenidas y su respectiva concordancia donde el 100 % del exoma esta representado en las variaciones encontradas mientras 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 cuando se comparan dos archivos de distintas variaciones, este puede ser abierto como un archivo de texto o cargado directamente en R utilizando la librería *gslib* 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

Cuadro **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 se 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

Cuadro **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 [51] 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.

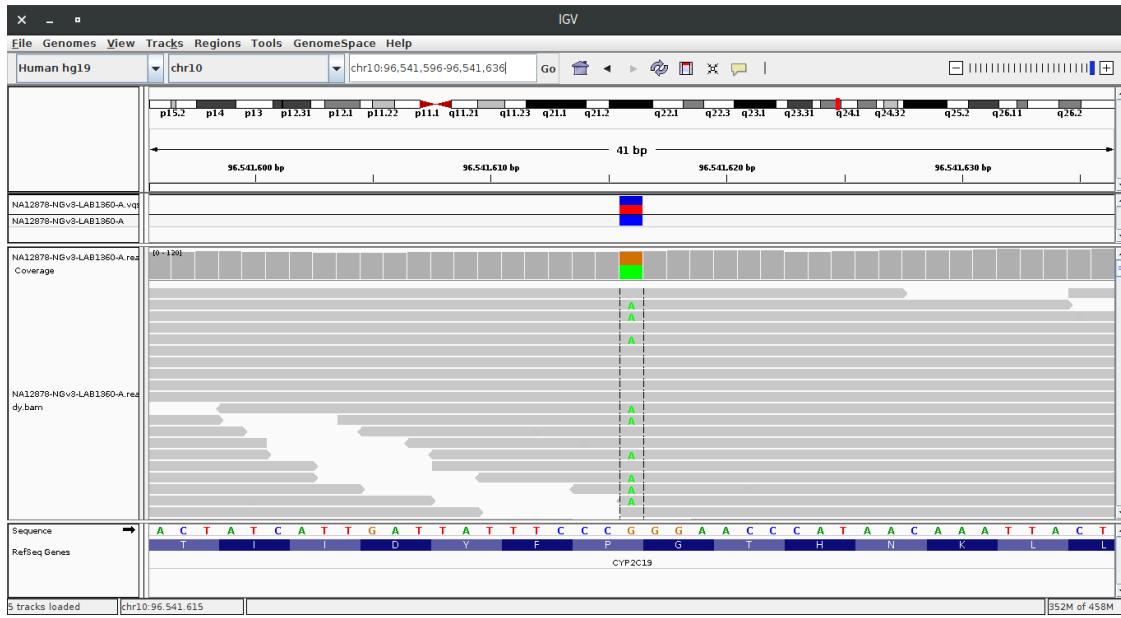


Figura 3.13: Imagen de la variante presente en el exoma público

3.5. Discusión

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 [50].

El presente conjunto de secuencias FASTQ se encuentra con buenos parámetros de calidad, aunque algunos módulos 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 mayoría de las lecturas obtenidas en la muestra y que posiblemente son secuencias duplicadas que no pertenecen al conjunto de secuencias real, a pesar 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 [50][52].

Además existen otras características que pueden generar impactos negativos dentro del análisis de datos de NGS divididas en dos grupos [53]:

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 [53].
2. Contaminación de las lecturas de especies conocidas o no conocidas en la secuencia objetivo, este error es frecuente y puede ser causado por un experimento artificial durante la preparación de la muestra, la construcción de la librería 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 experimental y por lo tanto si se cree que hay una contaminación lo ideal es realizar un trimming de las secuencias para remover la contaminación. **Nota:** Siempre y cuando estén en una baja proporción [53].

Las secuencias que se observan pueden ser duplicados de PCR que son un problema crítico cuando los fragmentos están sobre amplificados durante la preparación de las librerías, estos duplicados pueden aumentar a frecuencia alelica e incluir una detección errónea de variantes, esto es muy común en los datos de metagenómica, pero en nuestro caso los datos no son datos de metaagenómica si no de un solo individuo llama la atención de que solo estén al inicio de las lecturas y que el final de las lecturas esté adecuado esto podría indicar que más que un duplicado de PCR pueda ser un error de secuenciación al inicio de cada nuevo ciclo.[54].

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 óptima de los datos.

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 recalibration (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 [47].

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 3.7) 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 [47].

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 3.8), 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 HaplotypeCaller que mejora el llamado de variantes, y corrige algunas inconsistencias para la identificación de indels [55]. 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 [47].

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 mayoría 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 respecto a la identificación de SNPs (BWA-men) y HaplotypeCaller para la identificación de indels [49].

Variantes con un exoma NA12878.

Para este estudio se utilizo una muestra del genoma completo de la muestra NA12878 son de 34,886 variaciones [49] 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 [56]

Las dos resultados presentan una distribución similar en cuanto a las variantes por cromosoma y no hay variantes desconocidas dentro de la muestra, esto se debe a la alta curación que tiene este exoma, la figura 3.11 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 [47]. Esto se puede corregir por dos vias, aplicando un filtro de Quality by Depth (QD) ≥ 4 and Fisher Strand Bias (FS) $= \pm 30$ para dar un balance a la sensibilidad y la especificidad [57] 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 discovery 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 [58]. 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.

3.6. 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 parámetros 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.

Resumen

Se realizo la implementación y validación de un pipeline para la identificación variantes a partir de secuencias de exomicas a partir de muestras de pacientes colombianos y del genoma público de la muestra NA12878 donde se identificaron las variantes que presentes en el mismo, teniendo en cuenta las buenas practicas para el análisis de variantes lo que permitió desarrollar un mecanismo para obtener variantes de buena calidad.

4 Modelo de integración de la información

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, 59].

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 [59, 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 [60].

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 la información genómica y clínica son datos altamente heterogéneos con respecto a los modelos de datos que emplean normalmente, los esquemas de datos que especifican, los lenguajes de consulta que soportan y las terminologías que reconocen [61].

Para el caso de las variantes se tiene que para cada individuo hay 3.5 millones de variantes por individuo, estas variantes son almacenadas normalmente en el formato VCF (Formato de llamado de variantes), a su vez estos archivos pueden contener varias gigas de información, especialmente las muestras de genomas completos, que representan un problema para el almacenamiento dentro de las bases de datos, bases de datos que han sido desarrolladas para dar soluciones dependientes de las diferentes características y necesidades de los

laboratorios[20].

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. <https://www.djangoproject.com/>.

Este capítulo esta organizado en 1. Diseño e implementación de datos. 2. Gestión de datos clínicos y genómicos. 3. Conclusiones. 4. Resumen.

4.1. Diseño e implementación del modelo de datos

A continuación proponemos la utilización de una base de datos con información clínica y las variantes obtenidas a partir del pipeline. La figura 4.1 representa el esquema de datos que fue utilizado para realizar la integración de la información dentro de la base de datos.



Figura 4.1: Esquema de datos integrados

Teniendo en cuenta la información a utilizar se diseña el esquema EER que muestra la figura 4.2 con las tablas generadas por la aplicación para crear la base de datos propias de Django y las tablas de para la gestión de las variantes junto con la historia clínica.

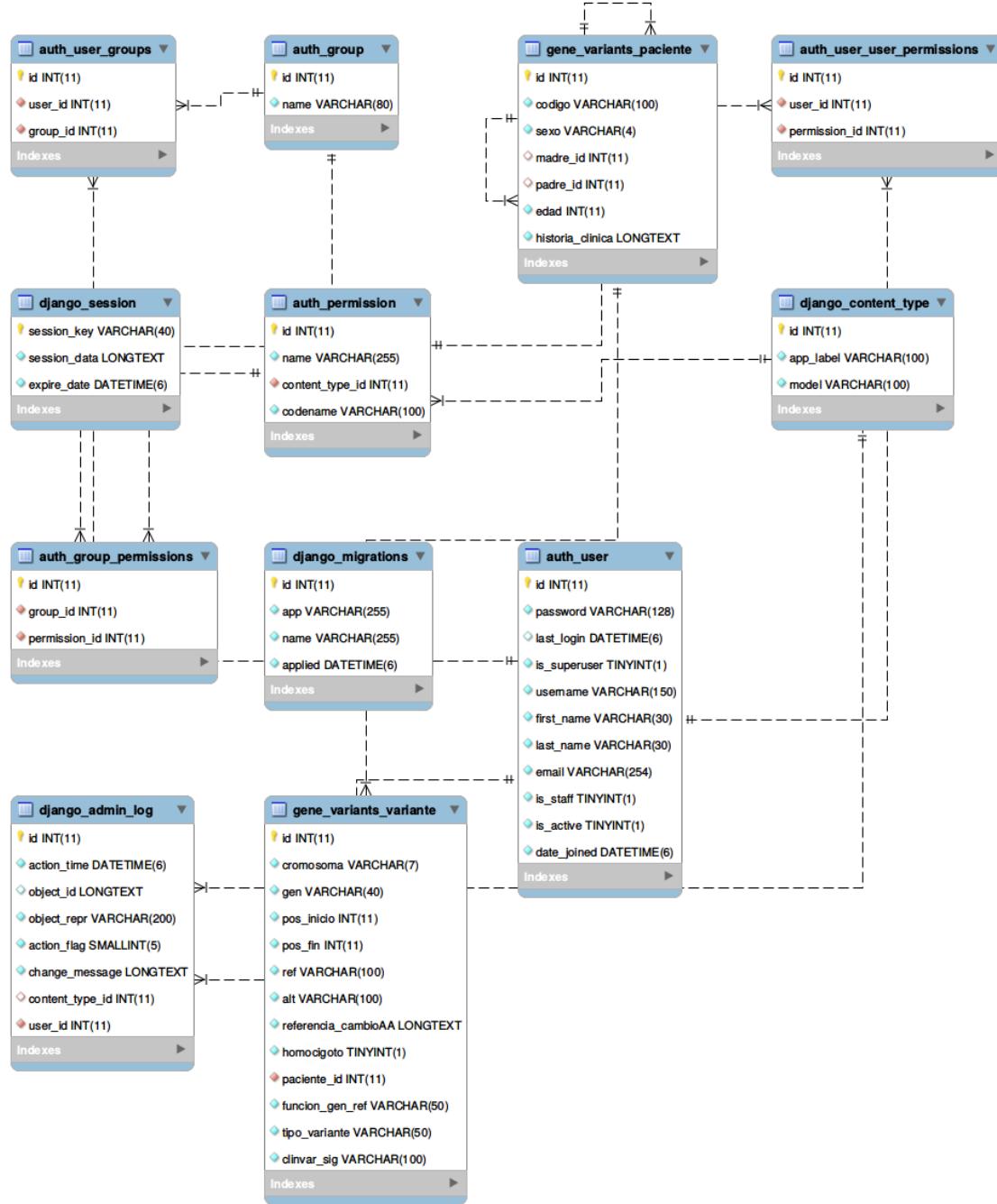


Figura 4.2: Modelo entidad relación

Las tablas diseñadas para gestionar las variantes y las historias clínicas son `gene_variants_paciente` que contienen:

- Edad: 0-99. Los recién nacidos o menores de un año tienen una edad de 0.
- Sexo: F o M según corresponda.

- Descripción: Que corresponde a la información clínica disponible.

Las historias clínicas fueron transcritas manualmente y cargadas desde un archivo de texto plano con un formato específico.

Las variantes con su historia clínica fueron cargadas mediante un script en bash disponible en <https://github.com/jevezelzse/variantesBD/blob/master/carga.bash>, donde se toman los archivos .csv de annovar junto con los archivos de texto que tienen la información clínica del paciente distribuida de la siguiente forma:

4.2. Gestión de datos genómicos y clínicos

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 4.3: Interfaz de ingreso para administrar la base de datos.

Inicialmente la figura 4.3, muestra la solicitud de usuario y contraseña para acceder a la aplicación, es diferente a la base de MySQL, pero puede tener una contraseña igual o diferente a la de la base de datos.

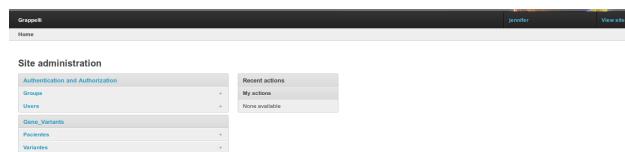


Figura 4.4: Interfaz de administración.

La figura 4.4, 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, 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, lo que permite actualizar los cambios que se reporten para la variante, por ejemplo variantes que por su alta frecuencia poblacional dejan de ser variantes y se convierten en referencias.

The screenshot shows the 'Add paciente' (Add patient) form in the Grappelli admin interface. It has fields for 'Name' (with a dropdown arrow), 'Sex' (with a dropdown arrow), 'Age' (text input), and 'Clinical history' (text area). At the bottom, there are buttons for 'Save and continue editing', 'Save and add another', and 'Save'.

Figura 4.5: Ingreso de pacientes.

En la figura 4.5 se muestra el formulario para ingresar una nueva historia o de modificar una historia clínica de un paciente de manera manual.

The screenshot shows a search results page for variants of the BRCA1 gene. The table has columns: ID, Chromosome, Gene, Pos. inicio, Pos. fin, Ref., Alt., Type-variant, Function, gen.ref., Phenotype, ClinVar, Ref. comp., and Annotations. There are 14 results listed, each with a checkbox and a link to the variant details. The annotations column shows various clinical variant descriptions.

Figura 4.6: Consulta a variantes

La figura 4.6 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 pacientes 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.

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 [60].

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 específico 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. [61].

4.3. Conclusión

La utilización de aplicaciones en Django permite que un bioinformático diseñe e implementar bases de datos aplicadas al diagnóstico 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 implementada la base de datos también es posible aplicar técnicas de minería de datos para optimizar los análisis de la información.

Resumen

En este capítulo se presentó el proceso de diseñar e implementar un sistema de información para la gestión de datos clínicos y genómicos, dada la importancia de tener toda la información integrada para hacer futuros análisis. Se utilizó la herramienta de Django como gestor de la base de datos, se transcribió las historias clínicas y se cargaron las variantes obtenidas para cada paciente, como resultado se generó un sistema de información que permite realizar consultas de variantes con las características clínicas de los pacientes.

5 Modelo de minería de datos

La necesidad de comprender los procesos biológicos que están implicados en las distintas enfermedades, a partir de la gran cantidad de datos biológicos que hay disponibles como las secuencias genómicas, los microarreglos, las interacciones proteicas, las imágenes biomédicas entre otros. Además la rápida adopción de las historias clínicas electrónicas proporciona una oportunidad de realizar investigaciones a gran escala. Por lo tanto las técnicas de minería de datos para el descubrimiento de conocimiento a partir de la obtención de información proveniente de diferentes fuentes son cada vez mas importantes en la investigación biológica y médica [62].

El mayor reto de la minería de datos genómicos esta en la extracción de información relevante de grandes volúmenes de datos clínicos y transformarlos en conocimiento, los mayores retos están en: a) La recolección de los datos clínicos y genómicos, b) recuperación de información relevante de datos y c) extracción de nuevos conocimientos de la información [44].

Este capitulo esta organizado en análisis exploratorio de los datos que se describe en la sección 5.1, el siguiente, es el análisis textual de información clínica que es discutido en la sección 5.2 en este apartado se describe el análisis de asociación de variantes con la información clínica. Finalmente, el apartado de 5.3 se presentan las conclusiones junto con el resumen del capítulo.

5.1. Análisis exploratorio de los datos.

Se realizó el análisis exploratorio de la información contenida dentro de la base de datos. Se tomó una muestra de 250 pacientes donados por el laboratorio Genetix S.A.S de los cuales solo 228 contaban con consentimiento informado para utilizar la información con fines de investigación.

Los datos fueron consultados desde la base de datos diseñada en el capítulo anterior y fue gestionada con la librería de python pandas [63]. Obteniéndose los siguientes resultados:

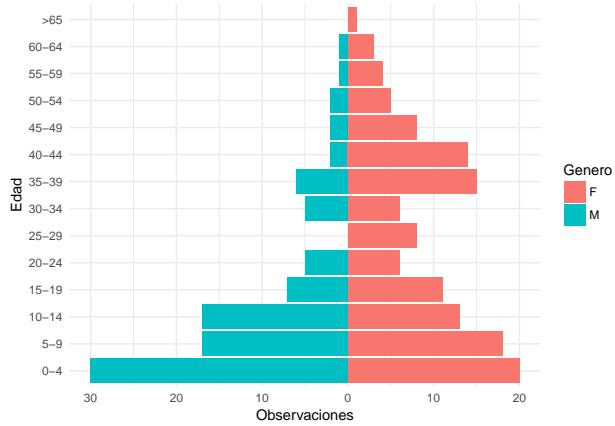
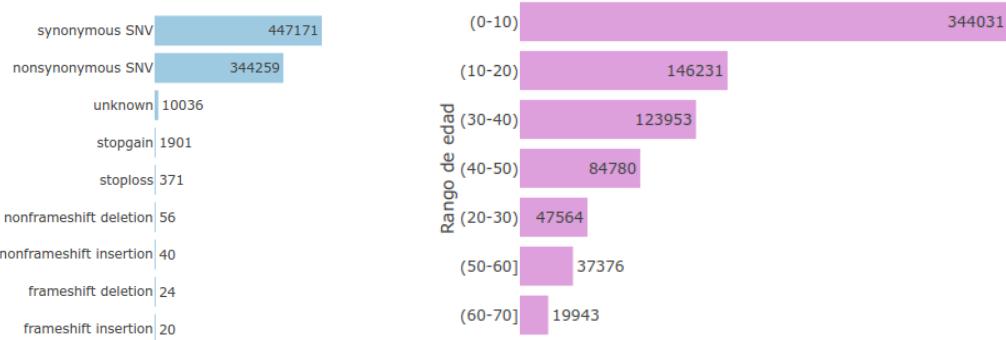


Figura 5.1: Distribución de rango de edades y géneros de los pacientes



(a) Distribución de variantes según su tipo.

(b) Distribución de variantes por rango de edad

Figura 5.2: Distribución del tipo de variantes

La base de datos contiene 228 pacientes de los cuales 133 son de género femenino y tienen un total de 468.485 variantes y 95 de género masculino con 345.239 de variantes obteniendo un total de 803.878 variantes. La figura 5.1 representa la distribución de pacientes por rango de edades y la figura 5.2 representa la distribución de variantes según su tipo. En la figura 5.2(a) muestra el número de variantes que son sinónimas y no sinónimas siendo las más frecuentes en la población, a nivel mundial se conoce que estos son los tipos de variantes más frecuentes[64].

Las variantes desconocidas son el tercer tipo de variante más frecuente dado que aún existe el problema de selección del transcripto para realizar la nomenclatura adecuada de las variantes, por lo que el anotador informa que son desconocidas [65]. La figura 5.2(b) muestra la distribución de las variantes identificadas según el rango de edad, siendo el rango con mayor

número de variantes los pacientes que se encuentran entre las edades de 0 a 10 años, dado a que es la población más representada dentro de la base de datos.

El estado alélico de las variantes (cigocidad) que se encuentran dentro de la base de datos se dividen en heterocigotas 458639 que corresponden al 57,05 % del total de las variantes y homocigotas 345239 que corresponden al 42,95 %. La distribución de la cigocidad de las variantes se puede explicar desde el error que se puede generar en la identificación de las variantes dado que durante el llamado de variantes es posible que una variante homocigota se catalogue como heterocigota o si durante el proceso de secuenciación se identifican erróneamente los nucleótidos [50][52].

5.2. Análisis textual de información clínica.

5.2.1. Preprocesamiento.

El proceso de limpieza y nacionalización de texto se realizo de la siguiente manera:

1. Remoción de stop words en español, tildes y caracteres especiales como la letra ñ y todos los documentos se unificaron en letras minúsculas.
2. Teniendo en cuenta la información clínica se creo un diccionario de sinónimos, donde se reemplazaron palabras que hacen referencia a una misma característica.
3. Calculo de la frecuencias de palabras dentro de los documentos.
4. Se removieron las palabras pam,pacientes, secuenciación y gen dado que no son un factor diferenciador de los documentos.

Resultados

Los resultados que se obtuvieron para la frecuencia de palabras fueron seno, cáncer, síndrome, sospecha y años. La figura 5.3 muestra las primeras 30 palabras más frecuentes y la nube de palabras de todos los documentos.

Las frecuencia de palabras nos muestra las principales características de la información clínica siendo las palabras cáncer y seno los principales fenotipos, también se encuentra la palabra síndrome que puede asociarse a diferentes enfermedades y la palabra sospecha hace referencia a diagnósticos ambiguos que pueden tener los pacientes, una de las contribuciones de la secuenciación es que basado en el fenotipo puede ayudar a un diagnóstico,entre diferentes síntomas y síndromes que pueden ser aplicados a enfermedades raras y complejas[66].

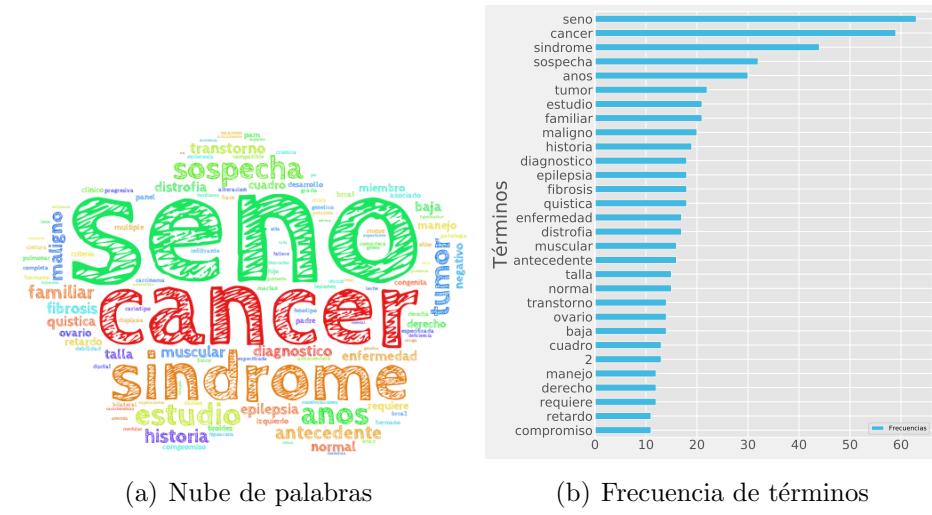


Figura 5.3: Frecuencias sin stop words y palabras sinónimas

5.2.2. Grupos de características clínicas.

En los procesos médicos, la relación entre los factores que pueden afectar la salud juega un papel importante. Una de las relaciones más comunes es la relación entre los genes y las enfermedades donde la secuenciación de exones tiene una alta aplicabilidad. Pero la identificación manual de este tipo de relaciones es compleja dada la cantidad de características que se pueden presentar como el diagnóstico propio de la enfermedad y/o la respuesta a los tratamientos [67].

La minería texto y puede ser aplicado al análisis en la medicina, donde el clustering (agrupamiento) puede ser considerado el método más importante que se utiliza en aprendizaje de maquina no supervisado que ha sido aplicado a diferentes problemas[67], teniendo en cuenta que no de los objetivos del agrupamiento de datos, es la identificación de grupos naturales en datos sin etiquetas[68].

Partiendo de lo anterior el presente trabajo se implemento un modelo de agrupamiento utilizando el kmeans para identificar grupos de características clínicas con la siguiente metodología:

1. Cálculo de la matriz tf-idf y se normalizo.
2. Estimación de el número de k optimo.
3. Implementación del algoritmo k-means.
4. Validación de los clusters.
5. Análisis de resultados.

Transformación de los datos.

El cálculo de la matriz tf-idf, se realiza a partir de frecuencia invertida con la ecuación

$$idf_i = \log_2 \frac{|D|}{|\{d \mid t_i \in d\}|}$$

siendo $|D|$ lo que denota el número total de documentos y donde $|\{d \mid t_i \in d\}|$ en que t_i aparece, la matriz de tf-idf es calculada a partir de la multiplicación de la frecuencia de términos y la frecuencia invertida $tf_{i,j} \cdot idf_i$ [69]. La figura 5.4 representa la matriz IDF-TF de las palabras que se encuentran dentro de la base de datos.

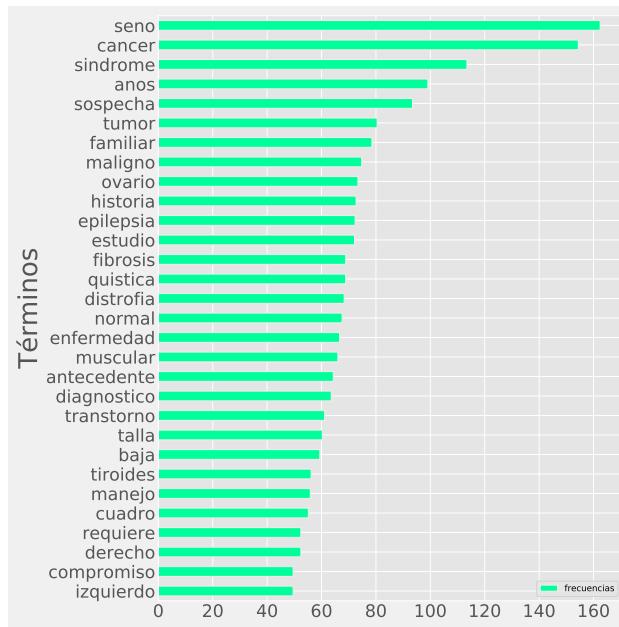


Figura 5.4: TF-IDF

Una vez obtenida la matriz tf-idf se normalizo y se le aplico la similaridad de coseno que es una de las medidas más populares para aplicar en documentos de texto. Esta medida tiene como ventaja que es independiente del largo del documento [70]. Esta similitud fue computada de acuerdo a la siguiente formula donde la similitud entre u y v es definida como según la librería scipy de python [71], Donde posteriormente los datos fueron utilizados para el clustering:

$$1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}.$$

donde $u \cdot v$ donde el punto es el producto de u y v .

Validación del modelo de clustering.

La selección del número óptimo de K se realizo utilizando el método del codo, este es uno de los métodos más antiguos para determinar el número de cluster, se realizan varios

experimentos iniciando por un $K = 2$ y realizando un incremento de 1, para los cuales se calcula el costo que conlleva cada una de las ejecuciones; entre más se aumente el número de K el costo disminuye y el número de K alcanza una meseta, este valor es el que se desea obtener, visualmente se realiza la identificación mediante un gráfico de error cuadrático y número de clusters , la razón es que al continuar el aumento del número de K los nuevos clusters son muy cercanos a otros ya generados [72].

El cálculo del error cuadrático vs el número de clusters se realizó utilizando la librería de python scikit learn, donde se computa el valor de la inercia que es calculada como la suma de cuadrados por cada punto cercano al centroide y es asignado al cluster. Así que $I = \sum_i(d(i, cr))$ donde cr es el centroide que fue asignado al cluster y d es la distancia cuadrada [73].

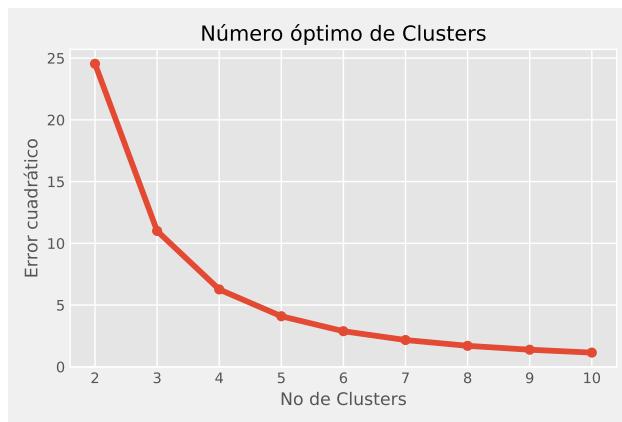


Figura 5.5: Número óptimo de clusters

Una vez se computó la inercia se realizó generó el gráfico del error cuadrático vs el número de clusters la figura 5.5 muestra el gráfico de codo obtenido, donde se puede seleccionar el cluster 5 y 6 como óptimo de K .

Para definir el número óptimo de K también se computó el coeficiente de Silhouette que es una evaluación de los clusters, donde los valores altos son relacionados a modelos que tienen clusteres bien definidos .El coeficiente está definido por cada muestra y está compuesta por dos valores que son [73, 74]:

- **a:** La distancia media entre una muestra y todos los puntos de la misma clase.
- **b:** La distancia media entre una muestra y todos los otros puntos en el próximo cluster más cercano.

El coeficiente Silhouette s para una sola muestra se da como:

$$s = \frac{b - a}{\max(a, b)}$$

Para un set de datos el coeficiente Silhouette es el promedio del coeficiente por cada muestra [73, 74]. En el presente trabajo los resultados del coeficiente de Silhouette fue de 0.534, adicionalmente se gráfico los valores de del coeficiente Silhouette para un $K = 5$ y se presenta en la figura 5.6:

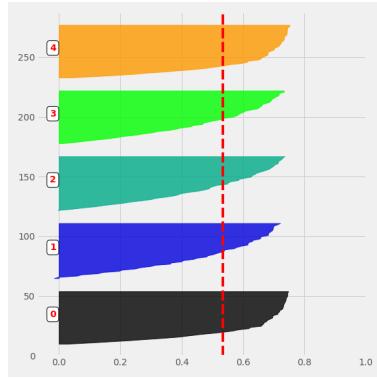


Figura 5.6: Valor Silhouette por cada cluster

Para los clusters obtenidos se calcularon las medidas de validación que están dentro de la librería scikit-learn [73] son:

- **Homogeneidad:** Definida como donde cada cluster contiene solo datos de una misma clase.
- **Integridad:** Donde todos los miembros de una misma clase son asignados al mismo clúster.
- **V-measure:** Es la medida armónica entre la homogeneidad y la integridad [75].

El cálculo de la homogeneidad y la integridad del cluster es realizada por:

$$h = 1 - \frac{H(C|K)}{H(C)}$$

$$c = 1 - \frac{H(K|C)}{H(K)}$$

donde $H(C|K)$ es la entropía condicional de las clases en cada asignación de cluster y que son calculadas por:

$$H(C|K) = - \sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \cdot \log \frac{n_{c,k}}{n_k}$$

y $H(C)$ es la entropia de clases y es calculada:

$$H(C) = - \sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log \frac{n_c}{n}$$

con n que es el número total de muestras n_c y n_k son el número de muestras asignadas respectivamente a la clase c y al cluster k , y finalmente $n_{c,k}$ son el número de muestras de las clases c asignadas al cluster k .

Finalmente el V-measure está definido de la siguiente manera [75]:

$$v = 2 \cdot \frac{h \cdot c}{h + c}$$

También se calcula Rand-Index que calcula una medida de similitud entre dos grupos al considerar todos los pares de las muestras y los pares de conteo que se asignan al mismo cluster o en diferentes grupos. Calculado de la siguiente manera [73]:

$$ARI = (RI - Expected_{RI}) / (max(RI) - Expected_{RI})$$

Los resultados de validación obtenidos fueron:

Para homogeneidad 0.296, para integridad 1.0, para el V-measure 0.457 y el Rand-Index fue 0. La homogeneidad perfecta sería con un valor de 1.0, en los presentes clusters presentan una baja homogeneidad, pero una integridad de 1.0 que significa que las etiquetas son perfectamente completas, esto se ve reflejado en el V-measure que es de 0.457 donde tenemos clusters con baja homogeneidad pero una alta integridad. Rand-Index se obtuvo un valor de 0.0 que muestra que las clases están separadas en diferentes clusters [73].

5.3. Asociación de grupos con variantes.

Una vez realizado el agrupamiento de la información clínica se aplicó un modelo de asociación de las variantes con los clusters obtenidos de la siguiente forma:

1. Consulta de las variantes que se encontraban en cada cluster.
2. Asociación de las variantes por cluster.
3. Asociación de las variantes por toda la información de la base de datos filtrada por el gen CFTR como caso de ejemplo.

La minería de datos frecuentes y las reglas de asociación es un método popular y bien investigado para describir las relaciones entre variantes en grandes bases de datos [76]. Las reglas de asociación (RA) muestran atributos con valores que ocurren frecuentemente en el set de datos, es posible obtener todos los posibles reglas de algunos atributos de acuerdo a la

presencia de otros atributos [77].

Las reglas de asociación se basan en un set de items(elementos) $I = \{i_1, i_2, \dots, i_n\}$ que son un conjunto de n atributos binarios. También se tiene que $D = \{t_1, t_2, \dots, t_m\}$ son el número de transacciones en la base de datos, cada transacción D tiene una identificación única y contienen un subconjunto de elementos en I . Una regla se define como una implicación de la forma $X \Rightarrow Y$ donde $X, Y \subseteq I$ y $X \cap Y = \emptyset$. Los sets de elementos son llamados *antecedentes y consecuentes* [76, 77]. La selección de reglas interesantes se realiza calculando la confianza y el soporte que son definidos como:

- Dados un set de datos $X \Rightarrow Y$, en una *regla de asociación* tiene una confianza c si c de nuestra transacción que contiene X pero que también contiene Y [78].
- Dados un set de datos $X \Rightarrow Y$ tiene una *regla de asociación* tiene un soporte s si $s\%$ de las transacciones en nuestra base de datos de transacciones que contienen $X \cup Y$ [78].
- Los algoritmos de asociación tratan de encontrar todas las reglas que tengan un mínimo de soporte y un mínimo de confianza[78].

5.3.1. Variantes vistas como transacciones.

Uno de los criterios más importantes para la clasificación de variantes es la frecuencia con la que se presentan las variantes dentro de una población según la asociación americana de genética médica [79], otro de los retos de los análisis de variantes es el estado alélico de las variantes, que se define como una forma alternativa de un mismo gen, en este caso se aplica a las variantes encontradas dentro de la secuenciación.

Este estado alélico puede ser de tres tipos, el homocigoto donde el individuo presenta la misma una sola variante (en términos al genoma de referencia), la otra forma son los heterocigotos, donde solo una variante es distinta con respecto a su referencia y los heterocigotos compuestos que son variantes heterocigotas pero que pueden estar dentro del mismo gen o asociadas con otros. Teniendo en cuenta lo anterior es importante visualizar el estado alélico de las variantes [14, 79] ya que pueden tener un impacto el fenotipo del paciente. La identificación entre la relación genotipo-fenotipo, se ingresan como frecuencia de variación, que para el trabajo caso serían las transacciones[80].

5.3.2. Experimentación

La confianza y el soporte para este trabajo se ajustó a partir de los resultados experimentales, donde se observó que el soporte es inversamente proporcional a la confianza, esto se debe a la cantidad de variantes que se encuentran dentro de la base de datos. Al correr un

experimento con soporte de 0.2 y una confianza de 0.9 no se generaban ningún tipo de regla, por lo tanto se fue disminuyendo en 0.1 el valor de la confianza y el soporte, finalmente se ajustó un soporte de 0.05 y una confianza de 0.6, dado que con un soporte de 0.1 solo se generaban 5 reglas, utilizando toda los datos disponibles. Una vez realizado este ajuste se dejaron los valores de soporte y confianza igual para todos los experimentos.

Un vez se obtuvieron los clusters de la base de datos, se realizaron 12 experimentos, los primeros 5 experimentos con todas las variante dentro del set de datos junto a su clusters, otro a todo el conjunto de datos aplicado con los datos y filtrado por el gen CFTR. Se volvieron a repetir los mismos experimentos pero removiendo las variantes sinónimas que son las más frecuentes dentro del conjunto de datos.

5.3.3. Resultados

Teniendo en cuenta las medidas de validación encontramos 5 clusters con las siguientes estructuras:

Cluster 1

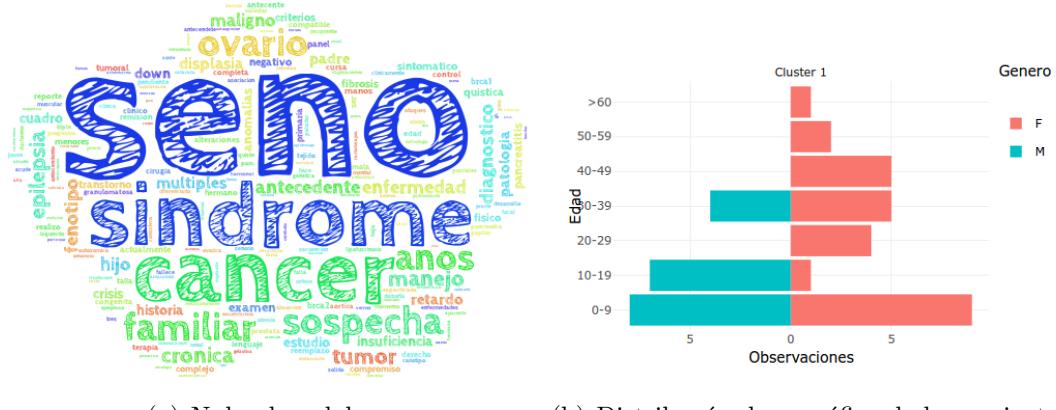


Figura 5.7: Cluster 1

La figura 5.7 representa el clúster 1 con la frecuencia de palabras que se agruparon para este clúster la figura 5.7(a) se muestra la frecuencia de palabras, siendo seno,síndrome y cáncer son las palabras más frecuentes, junto con ovario,familiar sospecha y epilepsia. La figura 5.7 representa la distribución de pacientes por edad y genero dentro del grupo por rango de edad en un intervalo de 10 años.

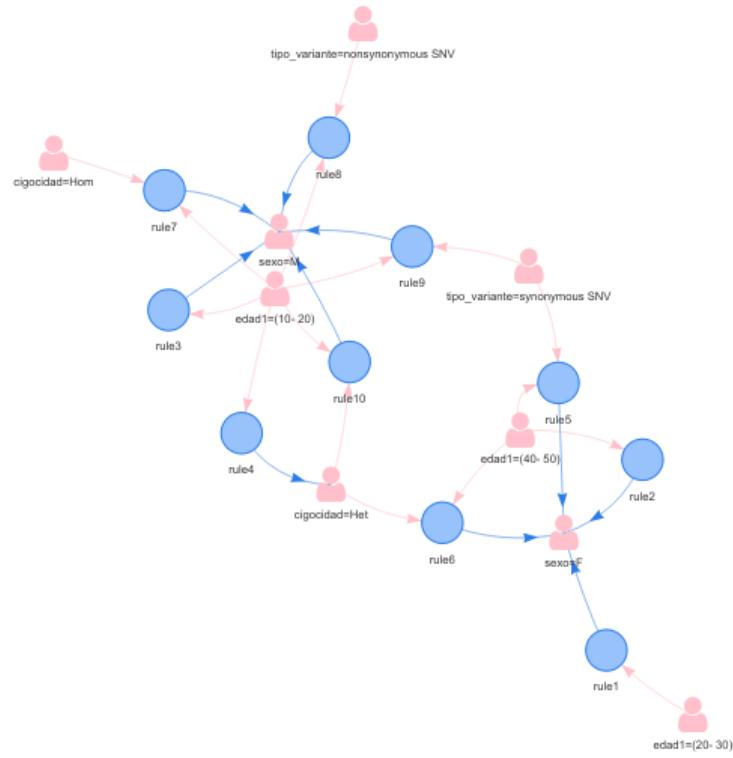


Figura 5.8: Reglas de asociación del cluster 1 con variantes sinónimas.

Las primeras 10 reglas obtenidas se representan mediante la figura 5.8 que muestra la asociación, sin remover las variantes sinónimas, el resultado obtenido muestra de dos tipos de variantes dentro del clúster 1. Para el genero masculino con variantes no sinónima y que son pacientes de edad entre 10 y 20 años, el estado alélico de las variantes es homocigoto, para este grupo se observa una alta diferencia en las reglas ambos géneros, donde las pacientes de genero femenino tienen variantes sinónimas con estado heterocigotas.

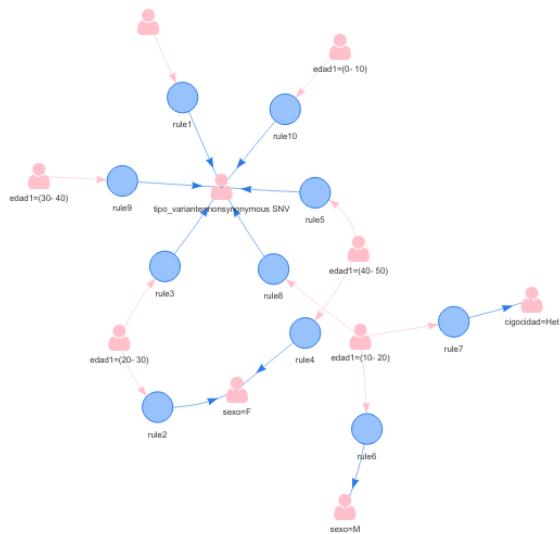


Figura 5.9: Reglas de asociación del cluster 1 sin variantes sinónimas.

Las 10 primeras reglas removiendo las variantes sinónimas que se muestran en la figura 5.9, nos muestra nuevamente la distribución de este tipo de variante dentro del grupo, donde los pacientes masculinos son pacientes entre 10 y 20 años, con variantes heterocigotas no sinónimas, mientras que para el genero femenino se tienen rangos de edad más amplios desde la edad de 0 a 50 años.

Cluster 2

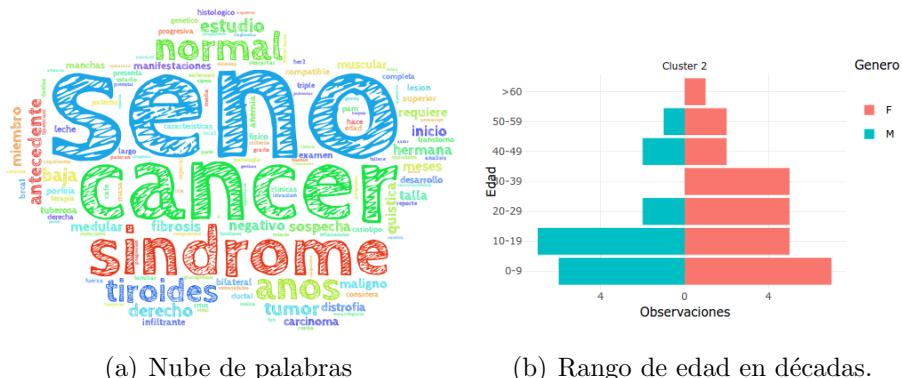


Figura 5.10: Cluster 2

En la figura 5.10(a) se observa que al igual que el cluster 1 las palabras más frecuentes son cáncer, seno y síndrome, pero aparecen palabras como antecedente tiroides y hermana, según la 5.10(b) se observa que rangos de edad entre 20 y 30 años, y mayores de 40 no hay pacientes masculinos, siendo este un cluster representado principalmente por pacientes femeninas.

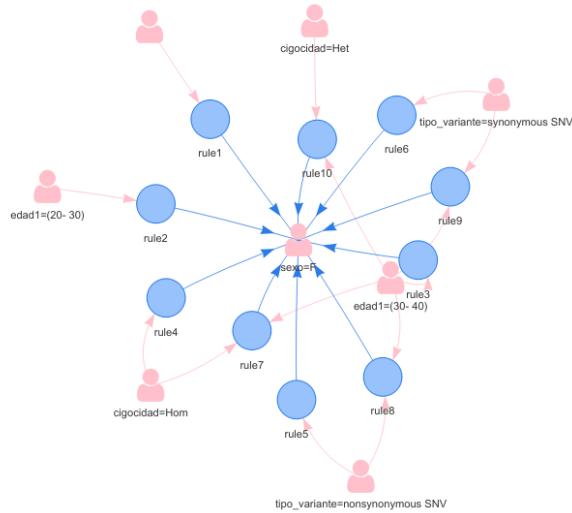


Figura 5.11: Reglas de asociación del cluster 2 con variantes sinónimas.

La figura 5.11, nos muestra únicamente una asociación de variantes al género femenino, que corresponde con la baja representación de pacientes de género masculino. La distribución del estado alélico homocigoto se presenta en mayor frecuencia con pacientes en edad de 30 y 40 años, y son variantes de tipo no sinónimo, a pesar de que el rango de edad no es el más representativo del cluster es el que presenta una mayor frecuencia de variantes.

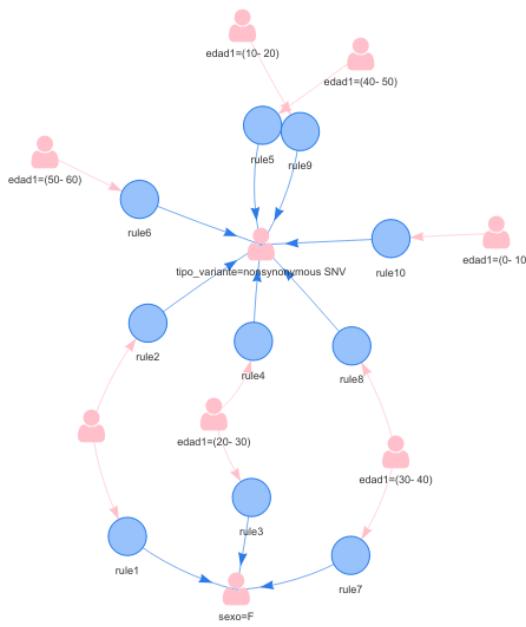


Figura 5.12: Reglas de asociación del cluster 2 con variantes sinónimas.

La figura 5.12, nos muestra la distribución de las variantes según los rangos de edad y que el tipo de variante no sinónimas, no muestra reglas para el estado alélico de las variantes dentro del clúster, solo las asociaciones entre rangos de edad y genero.

Cluster 3

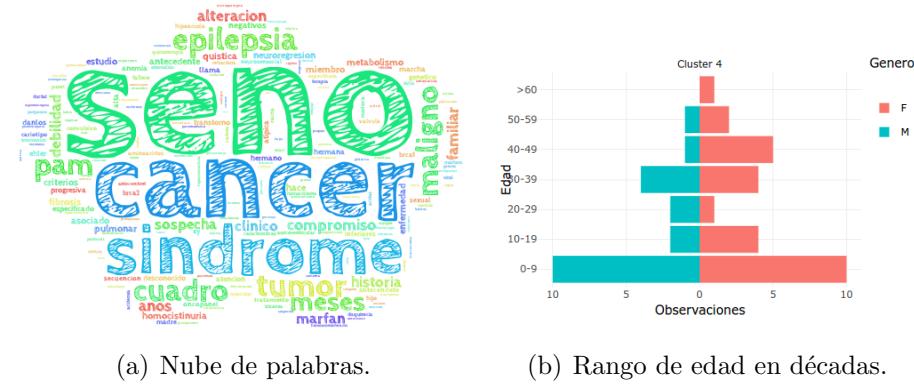


Figura 5.13: Cluster 3

La figura 5.13(a) nos muestra la frecuencia de palabras donde seno, cáncer y síndrome se tiene la palabra pam tumor maligno y epilepsia. La figura 5.13(b) muestra la distribución donde no hay pacientes de genero masculino para mayores de 60 años.

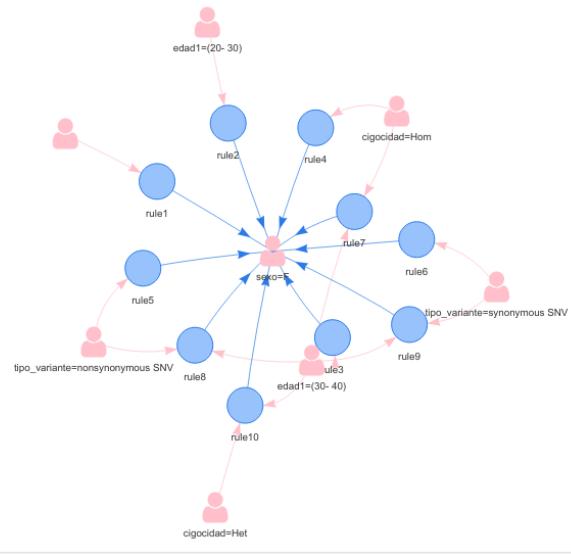


Figura 5.14: Reglas de asociación del cluster 3 con variantes sinónimas.

La figura 5.14 muestra las asociaciones entre las variantes sinónimas al género femenino, donde las variantes sinónimas se encuentran en mayor frecuencia a el rango de edad entre

30 y 40 y son heterocigotas, se presenta una frecuencia de pacientes entre los 20 y 30 años sin asociación a variantes o estado alélico.

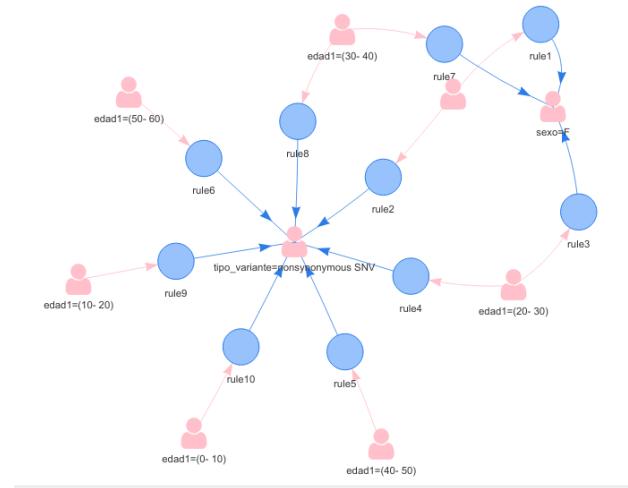


Figura 5.15: Reglas de asociación del cluster 3 sin variantes sinónimas.

La figura 5.15 muestra la asociación de rangos de edades con las variantes no sinónimas pero se encuentra que los rangos de edad de 30 y 40 y de 20 y 30 son de genero de género femenino, mientras que los demás rangos pertenecen no está asociados a un género en específico.

Cluster 4

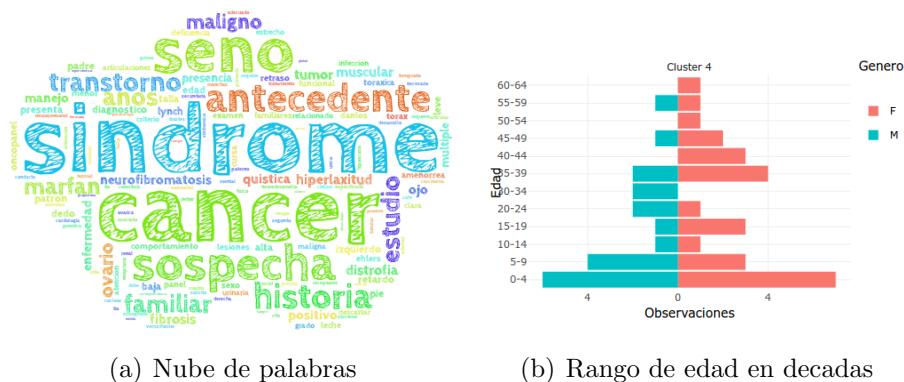


Figura 5.16: Cluster 4

La figura 5.16(a) muestra las frecuencias de palabras que son síndrome y cáncer, pero la palabra seno no es tan predominante como los clusters anteriores, se tienen otras palabras como sospecha, antecedente, historia y trastorno. La figura 5.16(b) muestra que para este cluster los rangos de edad de 40 a 45 años, de 50 a 55 años y mayores de 60 no cuentan con representación de pacientes de género femenino.

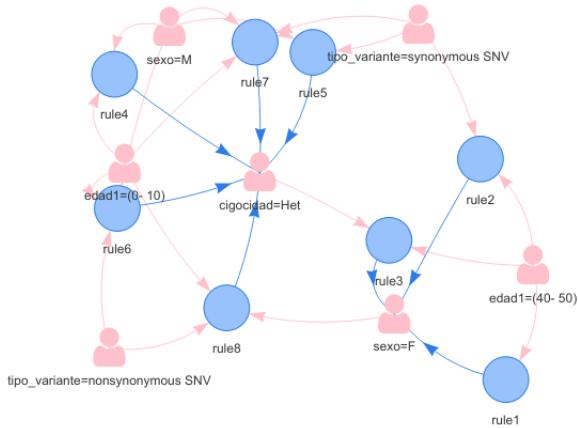


Figura 5.17: Reglas de asociación del cluster 4 con variantes sinónimas.

La figura 5.17 muestra la asociación de las variante heterocigotas a pacientes masculinos de tipo no sinónimas con un rango de edad de 0 a 10 años, mientras que las variantes sinónimas se asocian a pacientes con un rango de edad de 40 a 50 años y son pacientes femeninas.

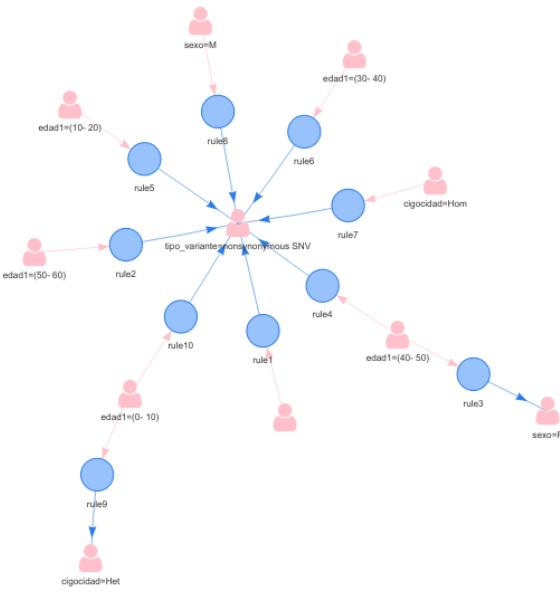


Figura 5.18: Reglas de asociación del cluster 4 sin variantes sinónimas.

La figura 5.18 muestra reglas donde las reglas del cluster nuevamente se discrimina que las variantes no sinónimas son femeninas y están en un rango de edad de (40-50), mientras que las variantes heterocigotas se encuentran en un rango de edad de 0 a 10 años.

Cluster 5

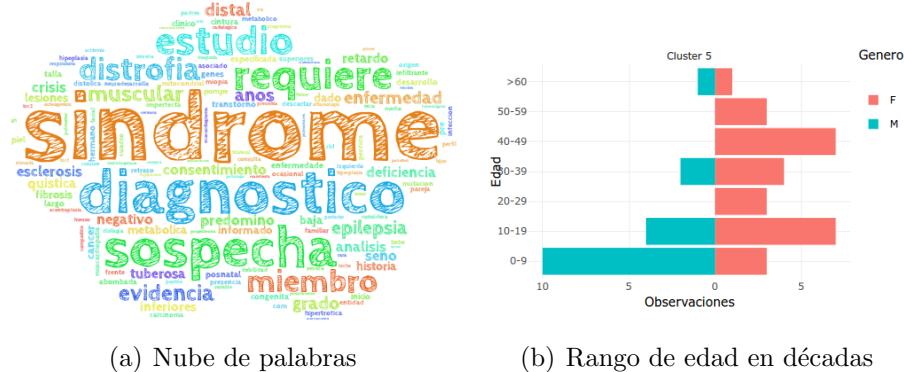


Figura 5.19: Cluster 5

La figura 5.19(a) presenta la frecuencia de palabras y este cluster a diferencia de todos los anteriores no presenta la palabras cáncer y seno, como las más frecuentes pero si presenta con más alta frecuencia son síndrome, diagnóstico, estudio, distrofia, requiere y miembro. La figura 5.19(b) muestra la distribución de pacientes por edad y género donde los rangos de 20 a 30 y 40 a 60 no se encuentran pacientes de género masculino, aunque tiene 10 pacientes masculinos en el rango de edad de 0 a 10 años.

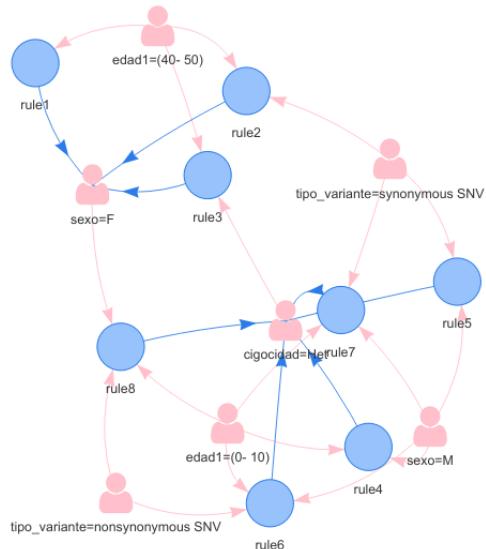


Figura 5.20: Reglas de asociación del cluster 5 con variantes sinónimas

La figura 5.20 muestra la asociación de las variantes al género femenino con un rango de edad de 40 a 50 años y de tipo sinónimas, mientras que las de género masculino a un rango de edad de 0 a 10 años con variantes no sinónimas.

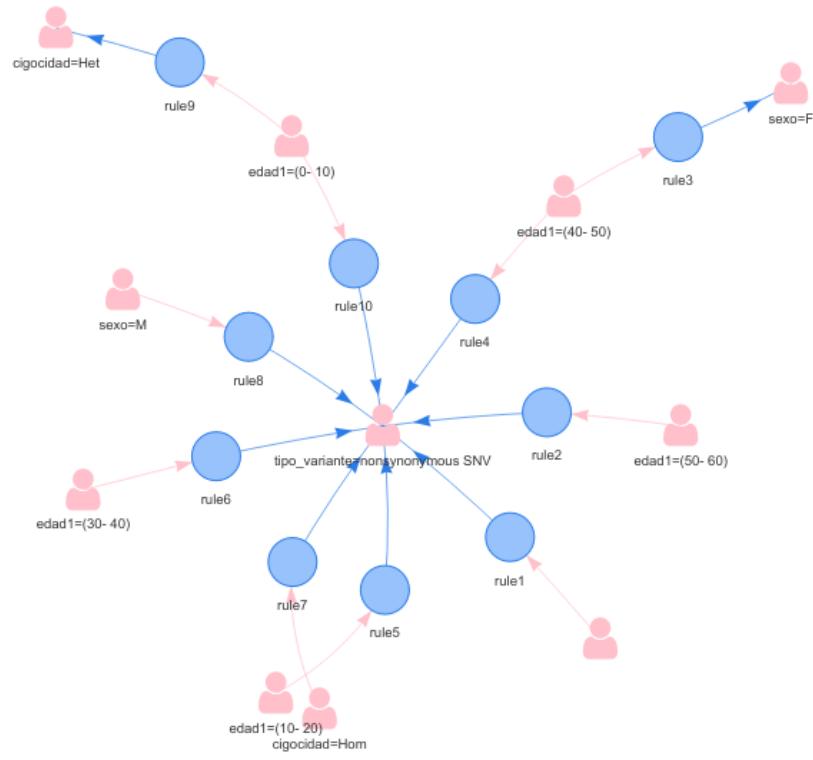


Figura 5.21: Reglas de asociación del cluster 5 sin variantes sinónimas

La figura 5.21 presenta que las variante no sinónimas son más frecuentes en los pacientes con un rango de edad de 0 a 10 años y que las variantes no sinónimas presentan la regla que las variantes se encuentran en un rango de 0 a 50 años de edad son de pacientes femeninas.

CFTR

Visualización de reglas de asociación para toda la base de datos utilizando el gen CFTR como filtro.

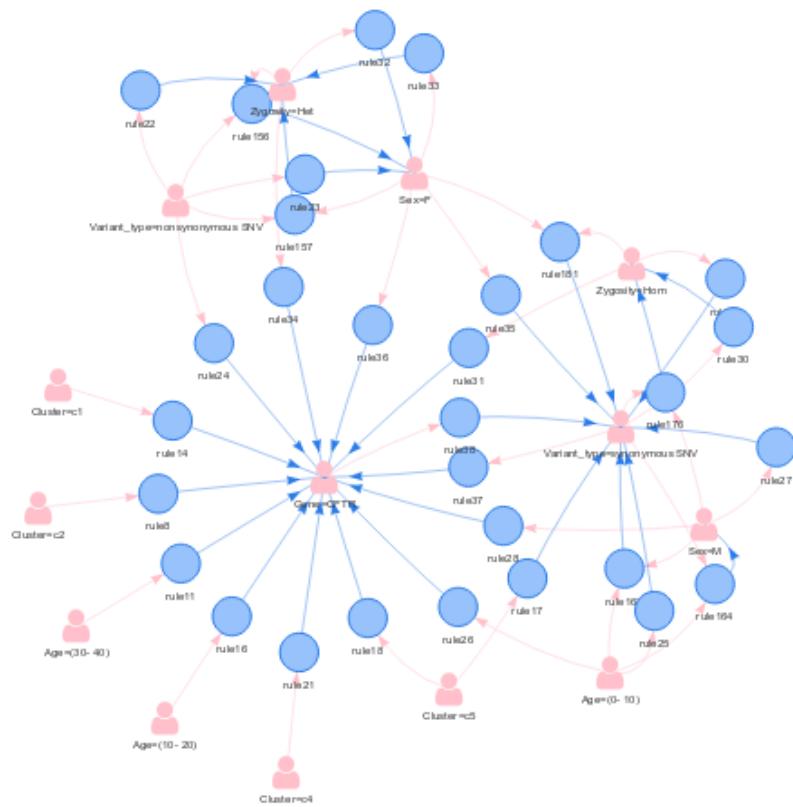


Figura 5.22: Reglas de asociación con variantes sinónimas

La figura 5.3.3 muestra las reglas 30 primeras reglas filtradas por el soporte con para toda la base de datos utilizando parámetro el gen CFTR, donde se observa que las variantes homocigotas son de tipo sinónimas y están más representadas en pacientes de ambos géneros, también se denota una frecuencia en pacientes que tienen entre 0 y 10 años de edad con variantes en este gen son de género masculino.

Las pacientes femeninas se tiene el caso de que las variantes son no sinónimas y no hay un rango de edad directamente asociado a las pacientes femeninas. Los rangos de edad de 10 a 20 y de 30 a 40 tienen una frecuencia de variantes para el gen CFTR igual o mayor al 60 %, lo mismo se presenta con cluster 1,2 y 4. Para cluster 5 si se presenta una alta frecuencia de variantes sinónimas, mientras que para el cluster 3 no aparecen reglas con alta frecuencia con variantes en el gen CFTR.

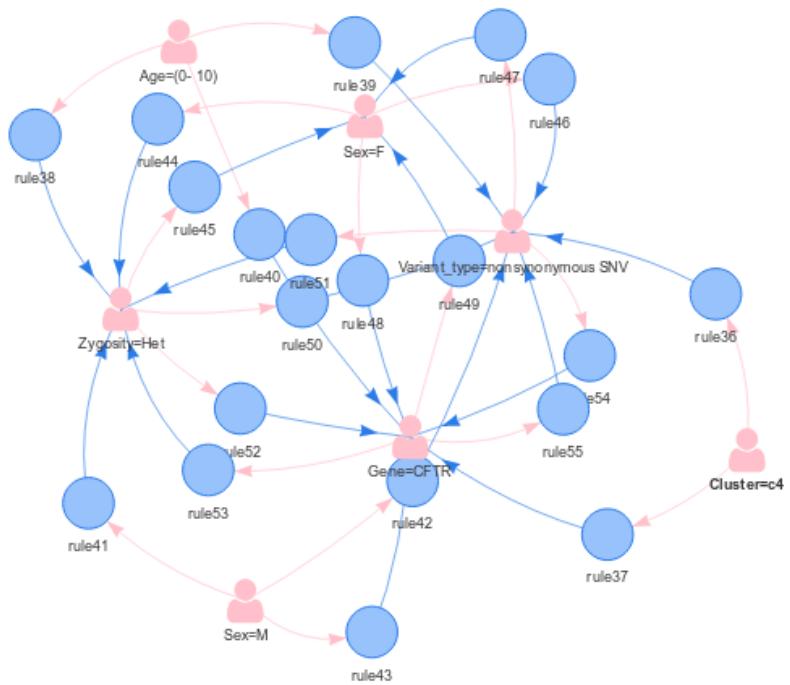


Figura 5.23: Reglas de asociación sin variantes sinónimas

La figura 5.23 muestra que el estado alélico de las variantes que no son sinónimas, y el estado alélico es heterocigoto para ambos generos, pero el rango de edad entre 0 y 10 años es frecuente para este tipo de variantes,a diferencia de la figura solo se observa una alta frecuencia de las variantes no sinónimas al gen CFTR en cluster 4.

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.

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. *Genomics, proteomics & bioinformatics*, 12(5):187–9, oct 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 Loguerchio, 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] Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke, and Zlatko Trajanoski. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, 15(2):256–278, 2014.
- [13] Merina Maharjan. Genome Analysis with MapReduce. pages 1–23, 2011.
- [14] Fady Hannah-Shmouni, Sara B. Seidelmann, Sandra Sirrs, Arya Mani, and Daniel Jacoby. The Genetic Challenges and Opportunities in Advanced Heart Failure. *Canadian Journal of Cardiology*, 31(11):1338–1350, 2015.
- [15] Brenton Louie, Peter Mork, Fernando Martin-Sánchez, Alon Halevy, and Peter Tarczy-Hornoch. Data integration and genomic medicine, feb 2007.
- [16] James D Watson and Francis H C Crick. Molecular structure of nucleic acids, 1953.
- [17] Angel Herráez. *Biología Molecular e Ingeniería Genética*. 2^a ed. Elsevier Ltd, Barcelona, 2012.
- [18] Pei Hui. Next Generation Sequencing: Chemistry, Technology and Applications. *Chemical Diagnostics Topics in Current Chemistry*, 336(1):1–18, 2012.
- [19] Jerzy K. Kulski. Next-Generation Sequencing — An Overview of the History, Tools, and “Omic” Applications. In *Next Generation Sequencing - Advances, Applications and Challenges*. InTech, jan 2016.
- [20] Joachim Kutzera and Patrick May. Data Integration in the Life Sciences. 6254:22–28, 2017.
- [21] Illumina. Whole Exome Sequencing — Detect exonic variants, 2017.
- [22] William S. Klug and Luis F. Pascual Calaforra. *Conceptos de genética*. Pearson Educacinón, 2013.
- [23] Eugenia Poliakov, David N Cooper, Elena I Stepchenkova, and Igor B Rogozin. Genetics in genomic era. *Genetics research international*, 2015:364960, 2015.

- [24] Xutao Deng. SeqGene: a comprehensive software solution for mining exome- and transcriptome- sequencing data. *BMC bioinformatics*, 12:267, jun 2011.
- [25] Aaron M Wenger, James T Robinson, Ahmet Zehir, and Jill P Mesirov. Variant Review with the Integrative Genomics Viewer. *Cancer research*, 77(21):31–35, 2017.
- [26] Jay Shendure. Human genomics: A deep dive into genetic variation. *Nature*, 536(7616):277–278, aug 2016.
- [27] T. Triplet and G. Butler. A review of genomic data warehousing systems. *Briefings in Bioinformatics*, 15(4):471–483, jul 2014.
- [28] Ira M. Lubin, Nazneen Aziz, Lawrence J. Babb, Dennis Ballinger, Himani Bisht, Deanna M. Church, Shaun Cordes, Karen Eilbeck, Fiona Hyland, Lisa Kalman, Melissa Landrum, Edward R. Lockhart, Donna Maglott, Gabor Marth, John D. Pfeifer, Heidi L. Rehm, Somak Roy, Zivana Tezak, Rebecca Truty, Mollie Ullman-Cullere, Karl V. Voelkerding, Elizabeth A. Worthey, Alexander W. Zarnek, and Justin M. Zook. Principles and Recommendations for Standardizing the Use of the Next-Generation Sequencing Variant File in Clinical Settings. *Journal of Molecular Diagnostics*, 19(3):417–426, 2017.
- [29] Gabriela Jurca, Omar Addam, Alper Aksac, Shang Gao, Tansel Özyer, Douglas Demetrick, and Reda Alhajj. Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends. *BMC Research Notes*, 9(1):236, 2016.
- [30] Vito Terlizzi, Giuseppe Castaldo, Donatello Salvatore, Marco Lucarelli, Valeria Raia, Adriano Angioni, Vincenzo Carnovale, Natalia Cirilli, Rosaria Casciaro, Carla Colombo, Antonella Miriam, Di Lullo, Ausilia Elce, Paola Iacotucci, Marika Comegna, Manuela Scorza, Vincenzina Lucidi, Anna Perfetti, Roberta Cimino, Serena Quattrucci, Manuela Seia, Federica Zarrilli, and Felice Amato. Genotype – phenotype correlation and functional studies in patients with cystic fibrosis bearing CFTR complex alleles. *Journal of Medical Genetics*, 54:224–235, 2017.
- [31] Somak Roy, Christopher Coldren, Arivarasan Karunamurthy, Nefize S. Kip, Eric W. Klee, Stephen E. Lincoln, Annette Leon, Mrudula Pullambhatla, Robyn L. Temple-Smolkin, Karl V. Voelkerding, Chen Wang, and Alexis B. Carter. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *Journal of Molecular Diagnostics*, 20(1):4–27, 2018.
- [32] Rayan Littlefield. An introduction into Data Mining in Bioinformatics.
- [33] David B. Searls. The Roots of Bioinformatics. *PLoS Computational Biology*, 6(6):e1000809, jun 2010.

- [34] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data."on cloud computing: Review and open research issues. *Information Systems*, 47:98–115, jan 2015.
- [35] Guomin Ren and Roman Krawetz. Applying computation biology and "big data"to develop multiplex diagnostics for complex chronic diseases such as osteoarthritis. *Bio-markers : biochemical indicators of exposure, response, and susceptibility to chemicals*, 20(8):533–9, 2015.
- [36] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigelski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, jan 2001.
- [37] Andrew Yates, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E Hunt, Sophie H Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J Martin, Thomas Maurel, William McLaren, Daniel N Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J Trevanion, Fiona Cunningham, Bronwen L Aken, Daniel R Zerbino, and Paul Flicek. Ensembl 2016. *Nucleic acids research*, 44(D1):D710–6, jan 2016.
- [38] Vincent Canuel, Bastien Rance, Paul Avillach, Patrice Degoulet, and Anita Burgun. Translational research platforms integrating clinical and omics data: A review of publicly available solutions. *Briefings in Bioinformatics*, 16(2):280–290, 2015.
- [39] Koichiro Higasa, Noriko Miyake, Jun Yoshimura, Kohji Okamura, Tetsuya Niihori, Hirotomo Saitsu, Koichiro Doi, Masakazu Shimizu, Kazuhiko Nakabayashi, Yoko Aoki, Yoshinori Tsurusaki, Shinichi Morishita, Takahisa Kawaguchi, Osuke Migita, Keiko Nakayama, Mitsuko Nakashima, Jun Mitsui, Maiko Narahara, Keiko Hayashi, Ryo Funayama, Daisuke Yamaguchi, Hiroyuki Ishiura, Wen-Ya Ko, Kenichiro Hata, Takeshi Nagashima, Ryo Yamada, Yoichi Matsubara, Akihiro Umezawa, Shoji Tsuji, Naomichi Matsumoto, and Fumihiko Matsuda. Human genetic variation database, a reference database of genetic variations in the Japanese population. *Journal of Human Genetics*, 61(6):547–553, 2016.
- [40] About the Human Variome Project: what we do and why we do it - Human Variome Project.

- [41] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. *Journal of Latex Class Files*, 13(9):1–20, 2014.
- [42] Ligia Bustos, Ricardo Moreno, and Nestor Duque. Modelo de una bodega de datos para el soporte a la investigación bioinformática. *Scientia*, XIII(037):13–18, 2007.
- [43] Stefan Naulaerts, Pieter Meysman, Wout Bittremieux, Trung Nghia Vu, Wim Vanden Berghe, Bart Goethals, and Kris Laukens. A primer to frequent itemset mining for bioinformatics. *Briefings in bioinformatics*, 36(2):3066–3076, 2013.
- [44] Dewan Md. Farid, Mohammad Abdullah Al-Mamun, Bernard Manderick, and Ann Nowe. An adaptive rule-based classifier for mining big biological data. *Expert Systems with Applications*, 64:305–316, 2016.
- [45] Mohammed J Zaki, George Karypis, and Jiong Yang. Data Mining in Bioinformatics (BIOKDD). *Algorithms for molecular biology : AMB*, 2:4, apr 2007.
- [46] 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.
- [47] 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 variant calls: the Genome Analysis Toolkit best practices pipeline*, volume 11. 2014.
- [48] 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.
- [49] 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.
- [50] Babraham Bioinformatics. FASTQC manual, 2016.
- [51] H Yang and K Wang. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc*, 10(10):1556–1566, 2015.
- [52] 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.

- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] Ellen Tsai, Rimma Shabatyan, 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.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] W Sujansky. Heterogeneous database integration in biomedicine. *Journal of biomedical informatics*, 34(2001):285–298, 2001.
- [62] Fei Wang, Xiao-li Li, Jason T L Wang, and See-kiong Ng. Guest Editorial: Special Section on Biological Data Mining and Its Applications in Healthcare. 14(3):501–502, 2017.
- [63] Wes McKinney. Pandas: a Foundational Python Library for Data Analysis and Statistics.

- [64] Wenqing Fu, Timothy D. O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M. Leal, Stacey Gabriel, David Altshuler, Jay Shendure, Deborah A. Nickerson, Michael J. Bamshad, and Joshua M. Akey. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2013.
- [65] Davis J. McCarthy, Peter Humburg, Alexander Kanapin, Manuel A. Rivas, Kyle Gaulton, Jean Baptiste Cazier, and Peter Donnelly. Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3), 2014.
- [66] Martine Tetreault, Eric Bareke, Javad Nadaf, Najmeh Alirezaie, and Jacek Majewski. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities, may 2015.
- [67] Koya Kawashima. Text Mining and Pattern Clustering for Relation Extraction of Breast Cancer and Related Genes. pages 1–5, 2017.
- [68] Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.
- [69] Gerard Salton Buckley and Christopher. Term-weighting approaches in automatic text retrieval. Information Processing and Management. *Information Processing and Management*, 24(5):513–523, 1988.
- [70] Anna Huang. Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, (April):49–56, 2008.
- [71] Eric Jones, Travis Oliphant, Pearu Peterson, and Others. SciPy: Open source scientific tools for Python.
- [72] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):2321–7782, 2013.
- [73] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [74] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C):53–65, 1987.

- [75] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language (EMNLP-ConLL'07)*, 1(June):410–420, 2007.
- [76] Michael Hahsler, Bettina Grün, and Kurt Hornik. Introduction to arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25, 2005.
- [77] Murat Karabatak and M. Cevdet Ince. An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2 PART 2):3465–3469, 2009.
- [78] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proc. 20th int. conf. very large data bases, VLDB*, (1215), pages 487–499, 1994.
- [79] Knight Diagnostic Laboratories, Medical Genetics, Oregon Health, Plank Road, Clinical Molecular, Nationwide Children, and Ohio State. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Medical Genetics*, 17(5):405–424, 2015.
- [80] René Breuer, Manuel Mattheisen, Josef Frank, Bertram Krumm, Jens Treutlein, Layla Kassem, Jana Strohmaier, Stefan Herms, Thomas W Mühleisen, Franziska Degenhardt, Sven Cichon, Markus Nöthen, George Karypis, Bipolar Disorder Genetics (BiGS) Consortium, Francis J McMahon, Marcella Rietschel, and Thomas G. Schulze. Genotype-phenotype association mining in bipolar disorder: market research meets complex genetics. *bioRxiv*, page 116624, mar 2017.