

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

Autor: Jennifer Vélez Segura

Director: Elizabeth León Guzmán

Asesor: Claudia Serrano

Grupo de Investigación – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

Mayo 2019



Contenido

1 Introducción

- Secuenciación
- Alineamiento
- Variantes
- Objetivos

2 Integración de datos

- Datos
- Identificación de variantes
- Modelo de datos

3 Minería de datos

- Preprocesamiento

4 Conclusiones

Contenido

1 Introducción

- Secuenciación
- Alineamiento
- Variantes
- Objetivos

2 Integración de datos

- Datos
- Identificación de variantes
- Modelo de datos

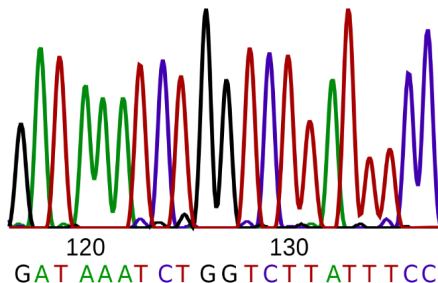
3 Minería de datos

- Preprocesamiento

4 Conclusiones

Secuenciación

La secuenciación es el proceso químico que permite identificar el orden de nucleótidos en molécula de ADN, ARN o una proteína [1].



Tomado de: <https://experiment.com/u/v0RZUQ>

Secuenciación

El proceso de secuenciación se realiza a través de los siguientes pasos:

- Extracción de ADN.
- Secuenciación de la muestra de ADN.
- Obtención de lecturas.
- Análisis de genes.
- Generación de reportes.



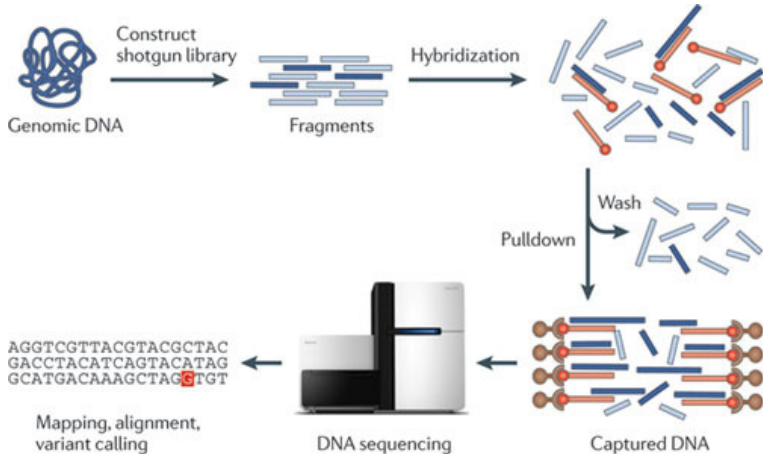
Secuenciación de siguiente generación (NGS)

Se han desarrollado diferentes metodologías para realizar secuenciación, estas son conocidas como secuenciación de alto rendimiento y tienen la capacidad de realizar secuenciaciones de manera masiva, rápida y económica [1].



<https://www.almacgroup.com/news/almac-group-to-collaborate-with-illumina-on-next-generation-sequencing-based-companion-diagnostic-development/>

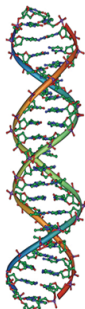
Secuenciación con Illumina



Alineamiento

Es el proceso de mapear las lecturas secuenciadas usando un genoma de referencia, que para humanos generalmente es el hg19.

Human genome



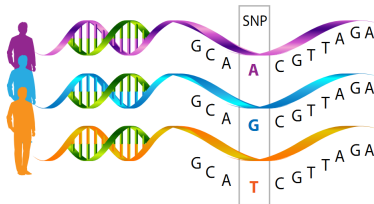
Short reads



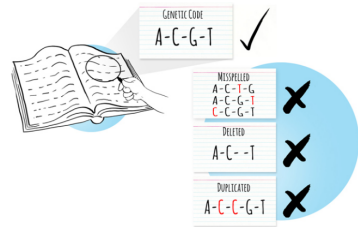
Alignment of reads to the reference genome and SNP calling



Variantes



(a) Variante de Nucleótido Simple SNV



(b) Variantes Indels

Figura 1: Tipos de variantes según el cambio de nucleótido

Variantes en el campo clínico

- 1 No existe un protocolo “gold standard” para realizar un análisis de secuencias.
- 2 Existe una gran cantidad de información obtenida que no es posible de analizar sin utilizar métodos computacionales.
- 3 La población colombiana no cuenta con estudios de este tipo y no se encuentra caracterizada como otras poblaciones.
- 4 En la mayoría de los estudios realizados no cuentan con la información clínica.



Objetivos

Proponer un modelo de minería de datos que permita asociar variantes con características clínicas de pacientes colombianos.

- Realizar la identificación de variantes en pacientes colombianos.
- Integrar la información clínica con las variantes obtenidas.
- Proponer un modelo de minería.
- Validar y visualizar los resultados del modelo de minería



Datos

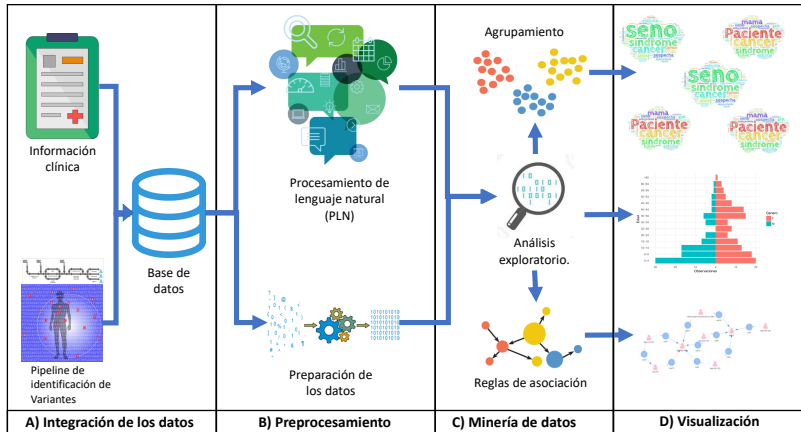
Origen de los datos

Los datos utilizados en el presente trabajo proviene de la información clínica de 228 pacientes colombianos, distribuidos en 133 pacientes femeninas y 95 pacientes masculino secuenciados en regiones codificantes de 4813 genes.

Estos datos fueron donados por el Laboratorio Genetix S.A.S.

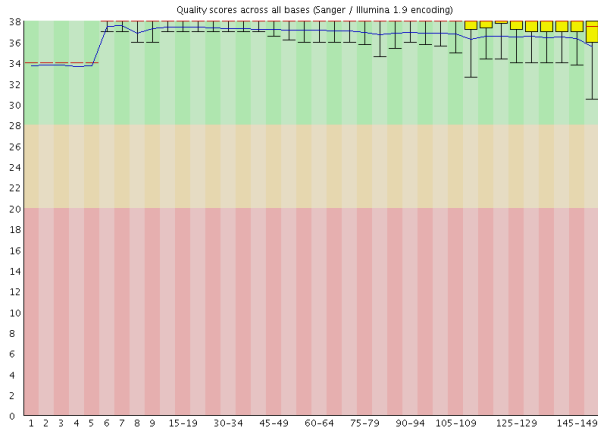


Flujo de trabajo



Calidad de la secuenciación

✓ Per base sequence quality



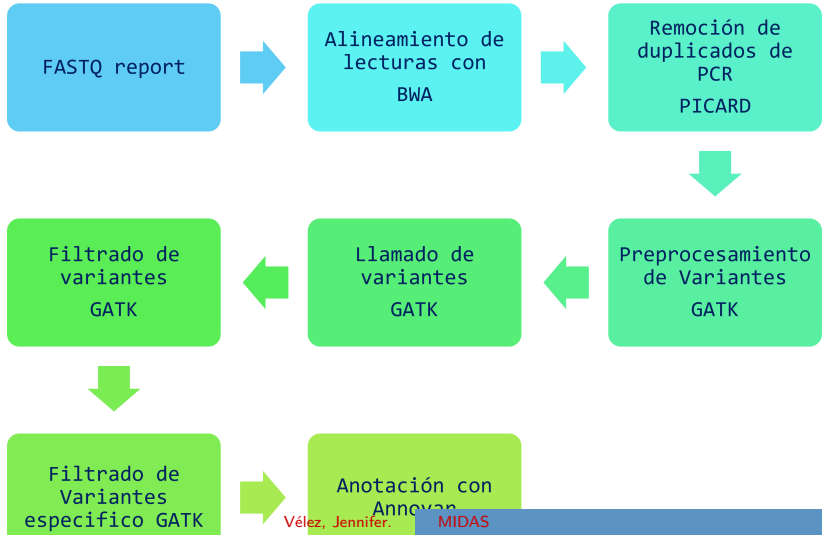
Identificación de variantes

Los pipelines bioinformáticos 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 [3]:

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.



Pipeline implementado en un cluster



Validación

1. Validación con datos internos. Datos de un paciente con 4813 genes(Illumina).
2. Validación con datos públicos. Datos del exoma completo NA12878.



Validación datos internos

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

Cuadro 1: Resultados de variantes obtenidas

Validación con datos internos

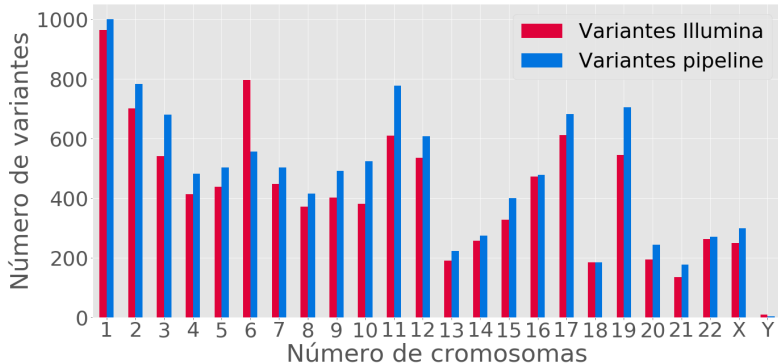


Figura 2: Distribución de variantes a lo largo de los cromosomas

Validación con el exoma NA12878

	Variantes Exoma		
	SNV	Indels	Total
Pipeline	30893	3324	34217
Públicas	29749	3101	32850

Cuadro 2: Tabla de Variantes obtenidas a partir de un exoma

Validación exoma NA12878

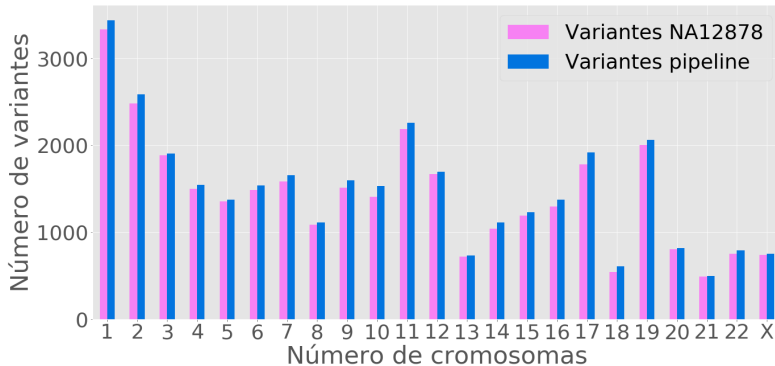


Figura 3: Distribución de variantes a lo largo de los cromosomas para los exomas

Validación exoma NA12878

TP	TN
32110	0
FP	FN
1033	0

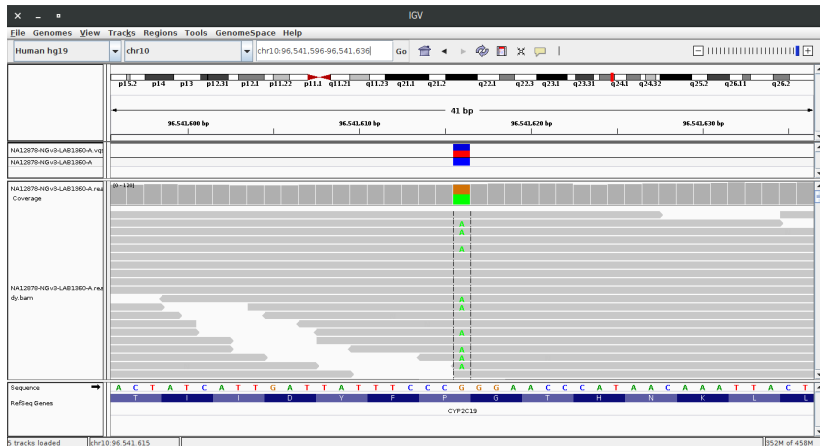
Cuadro 3: Matriz de confusión

Sensibilidad	Especificidad	PPV
96.88	100	100

Cuadro 4: Análisis de sensibilidad y especificidad

Variante en el Exoma NA1278

chr10,96541616,96541616,G,A,exonic,CYP2C19



Contenido

1 Introducción

- Secuenciación
- Alineamiento
- Variantes
- Objetivos

2 Integración de datos

- Datos
- Identificación de variantes
- Modelo de datos

3 Minería de datos

- Preprocesamiento

4 Conclusiones

Modelo de datos

- Autenticación de grupos.
- Autenticación de usuarios.
- Permisos de usuario.
- Migraciones de Django
- Administración de Django.
- Secciones de Django.

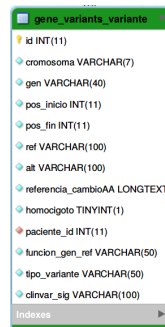


Modelo de datos



gene_variants_paciente	
id	INT(11)
codigo	VARCHAR(100)
sexo	VARCHAR(4)
madre_id	INT(11)
padre_id	INT(11)
edad	INT(11)
historia_clinica	LONGTEXT
Indexes	

(a) Datos de los pacientes



gene_variants_variante	
id	INT(11)
cromosoma	VARCHAR(7)
gen	VARCHAR(40)
pos_inicio	INT(11)
pos_fin	INT(11)
ref	VARCHAR(100)
alt	VARCHAR(100)
referencia_cambioAA	LONGTEXT
homocigoto	TINYINT(1)
paciente_id	INT(11)
funcion_gen_ref	VARCHAR(50)
tipo_variante	VARCHAR(50)
clinvar_slg	VARCHAR(100)
Indexes	

(b) Información de variantes

Figura 4: Gestión de variantes e información clínica

Interfaz de consulta

Grappelli jennifer [View site](#)

[Home](#) > [Gene_Variants](#) > [Variantes](#)

Variantes + Add variante

14 results 21352 total

<input type="checkbox"/>	Paciente	Cromosoma	Gen	Pos inicio	Pos fin	Ref	Alt	Tipo variante	Funcion gen ref	Homocigilo str	Clinvar stg	Ref. cambio
<input type="checkbox"/>	Paciente 33254 - F	chr17	BRCA1	41223084	41223094	T	C	nonsynonymous SNV	exonic	het	probable-non-pathogenic	BRCA1: exon14:c.A1525G;p.S509G BRCA1: exon14:c.A4696G;p.S1569G BRCA1: exon15:c.A1525G;p.S509G BRCA1: exon15:c.A4837G;p.S1613G BRCA1: exon16:c.A4900G;p.S1634G
<input type="checkbox"/>	Paciente 33254 - F	chr17	BRCA1	41234470	41234470	A	G	synonymous SNV	exonic	het	non-pathogenic	BRCA1: exon11:c.T4167C;p.S1389S BRCA1: exon11:c.T999C;p.S333S BRCA1: exon12:c.T4308C;p.S1436S BRCA1: exon12:c.T999C;p.S333S
<input type="checkbox"/>	Paciente 33254 - F	chr17	BRCA1	41244000	41244000	T	C	nonsynonymous SNV	exonic	het	1(thyroid)	BRCA1: exon10:c.A3548G;p.K1183R BRCA1: exon9:c.A3407G;p.K1136R
<input type="checkbox"/>	Paciente 33254 - F	chr17	BRCA1	41244435	41244435	T	C	nonsynonymous SNV	exonic	het	non-pathogenic	BRCA1: exon10:c.A3113G;p.E1036G BRCA1: exon9:c.A2972G;p.E991G
<input type="checkbox"/>	Paciente 33254 - F	chr17	BRCA1	41244936	41244936	G	A	nonsynonymous SNV	exonic	het	1(skin)	BRCA1: exon10:c.C2612T;p.P871L BRCA1: exon9:c.A2972G;p.E991G

0 of 14 selected

Figura 5: Ejemplo de consulta por gen

Contenido

1 Introducción

- Secuenciación
- Alineamiento
- Variantes
- Objetivos

2 Integración de datos

- Datos
- Identificación de variantes
- Modelo de datos

3 Minería de datos

- Preprocesamiento

4 Conclusiones

Preprocesamiento de la información clínica

1. Digitalización de la información clínica y carga a la base de datos.
2. Remoción de “stop words” en español, tildes y caracteres especiales como la letra ñ y todos los documentos se unificaron en letras minúsculas.
3. Creación de un diccionario de sinónimos, donde se reemplazaron palabras que hacen referencia a una misma característica, teniendo en cuenta la interpretación clínica.
4. Cálculo de la frecuencias de palabras dentro de los documentos.
5. Remoción de las palabras pam, pacientes, secuenciación y gen dado que no son un factor diferenciador de los documentos.



Preprocesamiento de lenguaje natural

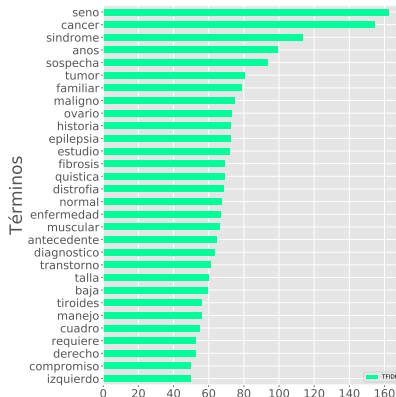
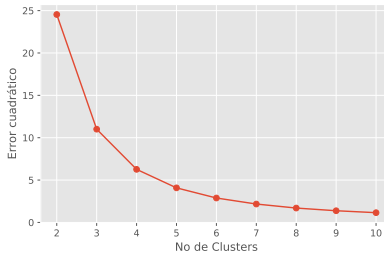
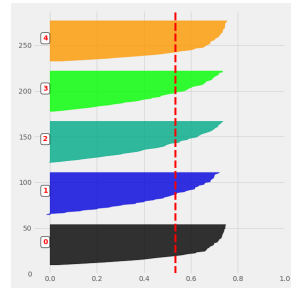


Figura 7: TF-IDF.

Agrupamiento de la información clínica



(a) Error cuadrático



(b) Valor Silhouette

Figura 8: Medidas de selección de número de grupos

Agrupamiento de la información clínica

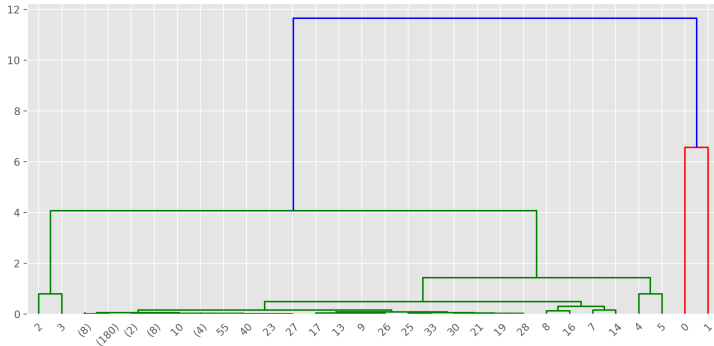


Figura 9: Agrupamiento jerárquico.

Variantes como transacciones

A partir de la base de datos se cálculo las variantes únicas obtenidas dentro del pipeline 29174 de la base se datos y se les asigno un ID a cada una.

Las variantes obtenidas se agruparon por gen y tipo de variante, rango de edad y genero del paciente.

		Items	Tipo de variante	Cigocidad	Rango de edad	Genero	Grupo
Transacciones	1	BRCA1	No sinónima	Het	(30-40)	F	C1
	2	RB1	Stop gain	Het	(0-10)	F	C5

Cuadro 5: Items y transacciones

Propuesta de visualización de resultados

Dashboard



Contenido

1 Introducción

- Secuenciación
- Alineamiento
- Variantes
- Objetivos

2 Integración de datos

- Datos
- Identificación de variantes
- Modelo de datos

3 Minería de datos

- Preprocesamiento

4 Conclusiones

Conclusiones

La utilización de la minería de datos en datos genómicos permite caracterizar e identificar variantes dentro de una población que no ha sido previamente caracterizada.

El modelo presenta la ventaja de poder utilizar información de diferentes fuentes para realizar un mejor análisis de datos genómicos.

- Este tipo de análisis son de bajo costo y eficientes.
- Puede ser aplicado a otros organismos según la disponibilidad de datos.



Trabajo futuro

- Aumentar la información como variantes de exomas y genomas, e información de regional de los pacientes para evaluar la población colombiana.
- Desarrollar una base de datos NoSQL para integrar la información procedente de diferentes fuentes y que sea permeable a cambios en el tamaño de la información.
- Modificar el cálculo de frecuencias del algoritmo Apriori de modo que permita, seleccionar items frecuentes, poco frecuentes e intermedios sin necesidad de que se calcule las asociaciones para todos los ítems.



Publicaciones



Figura 10: Presentación en eventos.

Agradecimientos

A mi familia que me acompaño en todo este proceso, a Sergio Solano y a Julián Cruz que me apoyaron con sus conocimientos, a la profesora Elizabeth León por dirigir este trabajo y a todo el equipo de Genetix SAS quienes donaron los datos utilizados en este trabajo.



GRACIAS!



Referencias I

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.

Thomas Triplet and Gregory Butler. A review of genomic data warehousing systems. *Briefings in Bioinformatics*, 15(4):471–483, jul 2014.



Referencias II

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.