Análisis de datos de Exomas

PEC2 Análisis de datos ómicos

Diego Vallarino
Junio, 2021

Contenido

1. Abstract

La secuenciación de próxima generación (NGS) está revolucionando rápidamente la forma en que se realiza la investigación sobre los determinantes genéticos de las enfermedades constitucionales. La técnica es muy eficaz con millones de lecturas de secuencia que se producen en un período de tiempo corto y a un costo relativamente bajo. Específicamente, la NGS dirigida es capaz de centrar las investigaciones en regiones genómicas de particular interés basadas en la enfermedad de estudio.

Esto no solo reduce aún más los costos y aumenta la velocidad del proceso, sino que también disminuye la carga computacional que a menudo acompaña a NGS. Aunque la NGS dirigida está restringida a determinadas regiones del genoma, lo que impide la identificación de posibles nuevos loci de interés, puede ser una técnica excelente cuando se enfrenta a una enfermedad fenotípica y genéticamente heterogénea, para la que existen asociaciones genéticas previamente conocidas.

Debido a la naturaleza compleja de la técnica de secuenciación, es importante adherirse estrictamente a los protocolos y metodologías para lograr lecturas de secuenciación de alta cobertura y calidad. Además, una vez que se obtienen las lecturas de secuenciación, se utiliza un sofisticado flujo de trabajo bioinformático para mapear con precisión las lecturas en un genoma de referencia, llamar variantes y garantizar que las variantes pasen las métricas de calidad.

2. Objetivo

El objetivo de este trabajo es llevar a cabo una búsqueda de variantes minoritarias en datos de exomas. En concreto nos planteamos encontrar variantes pequeñas (SNVs/indels) en DNA genómico humano usando un archivo que hemos obtenido del proyecto de los 1000 genomas a través del repositorio ISGR. En este caso se extrajo un conjunto aleatorio de un millón de secuencias ("shortreads") del cromosoma 22.

3. Materiales y Métodos

Un típico *workflow* para el descubrimiento de variaciones implica los siguientes pasos (por ejemplo, ver Nielsen et al. 2011):

a) Cargar los datos

El proceso de levantar los datos implicó seleccionar la muestra de datos que se proveyó como material del PAC2, y se cargó con las siguientes características según se muestra en la siguiente figura.

Figura 1: carga de datos utilizando el type como **fastq** para la carga de datos

b) Mapping reads contra el genoma de referencia y realizar un control de calidad de los datos

Este paso sirve para identificar posibles problemas con la secuencia sin procesar. Se ingresan los datos antes de embarcarse en cualquier paso de análisis "real".

Algunos de los problemas típicos con los datos NGS se pueden mitigar procesando previamente la secuenciación afectada reads antes de intentar mapearlos en el genoma de referencia. Detectar algunos otros problemas más graves desde el principio puede al menos ahorrarle mucho tiempo dedicado al análisis de datos de baja calidad que no vale la pena el esfuerzo.

Figura 2: utilización de **FastQC** para control de calidad

c) Establecer umbrales de conjuntos de datos de BAM, por ejemplo, reteniendo emparejados, correctamente mapeados reads

Se importaron en Galaxy Archivos FASTQ correspondientes a datos de extremo emparejado que podríamos obtener directamente de una instalación de secuenciación. Durante la secuenciación, se introducen errores, como la llamada incorrecta de nucleótidos. Los errores de secuenciación pueden sesgar el análisis y dar lugar a una mala interpretación de los datos. El primer paso para cualquier tipo de secuenciación de datos es siempre comprobar su calidad, actividad que se realizó en la etapa anterior.

Actualmente, hay más de 60 mapeadores diferentes y su número está creciendo. En este trabajo, usaremos Bowtie2 , una herramienta de código abierto rápida y eficiente en memoria, particularmente buena para alinear la secuenciación de los reads de aproximadamente 50 hasta miles de bases hasta genomas relativamente largos.

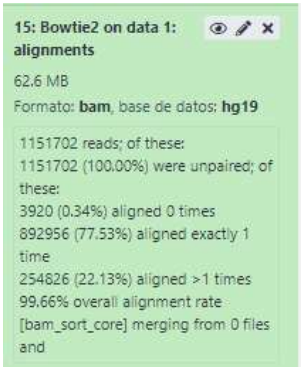

Figura 3: utilización de **Bowtie2** para control de calidad

d) Buscar diferencias entre las secuencias alineadas y el genoma de referencia

En esta etapa se utilizó FreeBayes la cual tiene una serie de características que simplifican los flujos de trabajo de descubrimiento de variantes. Estos incluyen:

- **La realineación de Indel se logra internaLLY** usando un método independiente-lectura, y los problemas resultantes de alineaciones discordantes se reducen drásticamente a través de la detección directa de haplotipos
- **La necesidad de recalibrar la calidad de la base se evita** mediante la detección directa de haplotipos. Los errores de la plataforma de secuenciación tienden a agruparse (por ejemplo, al final dereads) y generan haplotipos únicos que no se repiten en un locus dado.
- **La recalibración de la calidad variable se evita** incorporando una serie de métricas, como el sesgo de ubicación de lectura y el equilibrio de alelos, directamente en el modelo bayesiano.
- **Capacidad para incorporar casos no diploides**, como conjuntos de datos agrupados o datos de muestras poliploides.

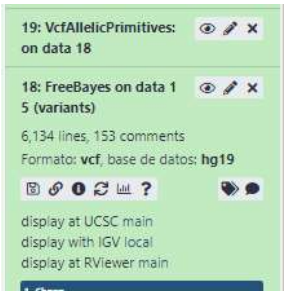

Figura 4: utilización de **FreeBayes** para diferencias de secuencias

e) Filtrar las variantes (por ejemplo, con SnpEff)

En este punto, estamos listos para comenzar a anotar variantes usando SnpEff " ... *anota y predice los efectos de las variantes en los genes (como los cambios de aminoácidos)* ... " y, por lo tanto, es fundamental para la interpretación funcional de los datos de variación.

SnpEff generará dos salidas:

1. un archivo VCF anotado
2. un informe HTML

El informe contiene una serie de métricas útiles, como la distribución de variantes entre características genéticas

Figura 5: utilización de **SnpEff** para filtrar variantes

4. Resultados

Las estadísticas del análisis muestran más de un millón de secuencias, tal cual se plantearon en la letra del PEC2.

Measure	Value
Filename	exerciseDataset_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1151702
Sequences flagged as poor quality	0
Sequence length	76
%GC	53

Figura 6: resumen estadístico

En la siguiente figura, para cada posición, se dibuja un diagrama de caja con:

- el valor mediano, representado por la línea roja central
- el rango intercuartil (25-75%), representado por el cuadro amarillo
- los valores de 10% y 90% en los bigotes superior e inferior
- la calidad media, representada por la línea azul

El eje *y* muestra los puntajes de calidad. Cuanto mayor sea la puntuación, mejor será la llamada base. El fondo del gráfico divide el eje *y* en puntuaciones de muy buena calidad (verde), puntuaciones de calidad razonable (naranja) y reads de mala calidad (rojo).

Es normal con todos los secuenciadores de Illumina que la puntuación de calidad media comience más baja en las primeras 5-7 bases y luego aumente. La calidad de readsen la mayoría de las plataformas caerá al final de la lectura. A menudo, esto se debe a la disminución de la señal o la fase durante la ejecución de secuenciación. Los recientes desarrollos en química aplicada a la secuenciación han mejorado esto un poco, pero reads ahora son más largos que nunca.

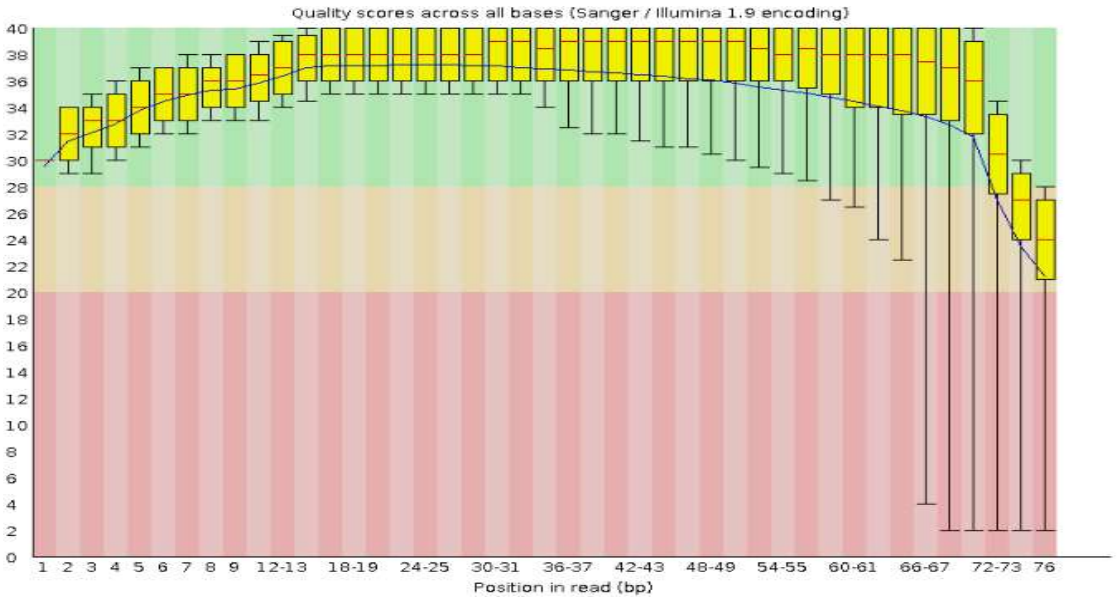

Figura 7: calidad de las bases/reads

A continuación, se muestran cómo ha sido la duplicación de los niveles, y como ha sido la distribución de los GC con respecto a la cueva teórica.

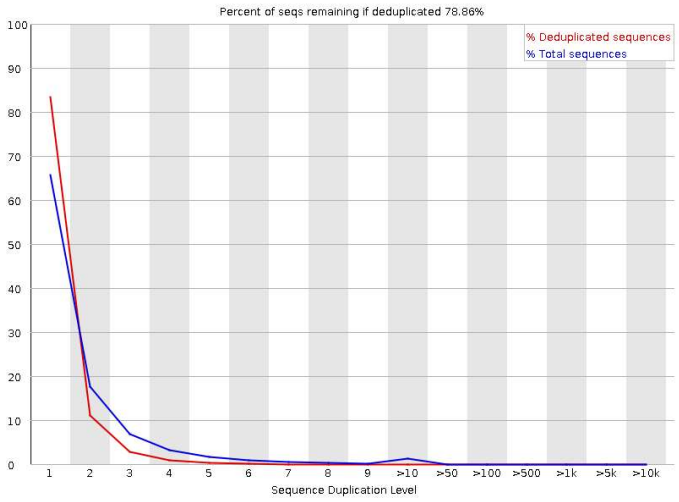

Figura 8: porcentaje de duplicación de las secuencias

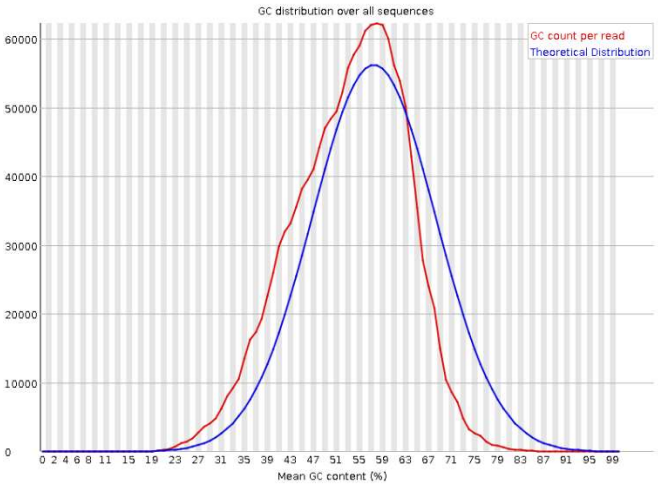

Figura 9: distribución de las secuencias en términos relativos con el modelo teórico

Podemos ver que el comportamiento de la secuencia sigue la curva teórica con un porcentaje de ajuste relevante.

Por último, se muestra la calidad de los scores por secuencia

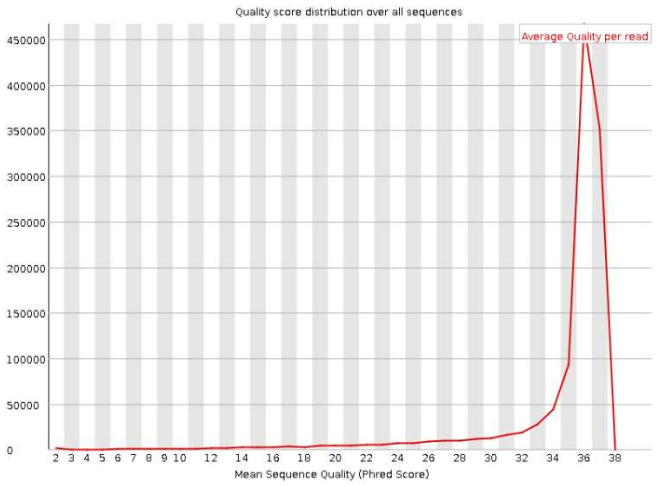

Figura 10: calidad del score de las frecuencias

A continuación, se presentan los efectos por tipo y por región. Particularmente se resalta la cantidad de intrones que se muestran en la estructura. Estas secuencias no contienen información propiamente dicha sobre los aminoácidos que darán lugar a la proteína, sin embargo, son importantes funcionalmente, puesto que en parte son responsables, por ejemplo, del barajado de exones. Los vertebrados superiores como el ser humano tienen un porcentaje muy elevado de intrones y secuencias que no codifican proteínas.

Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	330	1.833%
5_prime_UTR_premature_start_codon_gain_variant	14	0.078%
5_prime_UTR_variant	247	1.372%
conservative_inframe_deletion	1	0.006%
disruptive_inframe_deletion	5	0.028%
disruptive_inframe_insertion	1	0.006%
downstream_gene_variant	1,730	9.611%
frameshift_variant	4	0.022%
intergenic_region	1,668	9.266%
intron_variant	7,640	42.442%
missense_variant	2,297	12.76%
non_coding_transcript_exon_variant	520	2.889%
protein_protein_contact	2	0.011%
sequence_feature	207	1.15%
splice_acceptor_variant	15	0.083%
splice_donor_variant	82	0.456%
splice_region_variant	209	1.161%
start_lost	2	0.011%
stop_gained	18	0.1%
stop_lost	12	0.067%
structural_interaction_variant	805	4.472%
synonymous_variant	877	4.872%
upstream_gene_variant	1,315	7.305%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	1,730	9.775%
EXON	4,515	25.511%
INTERGENIC	1,668	9.425%
INTRON	7,385	41.728%
SPLICE_SITE_ACCEPTOR	15	0.085%
SPLICE_SITE_DONOR	82	0.463%
SPLICE_SITE_REGION	190	1.074%
TRANSCRIPT	207	1.17%
UPSTREAM	1,315	7.43%
UTR_3_PRIME	330	1.865%
UTR_5_PRIME	261	1.475%

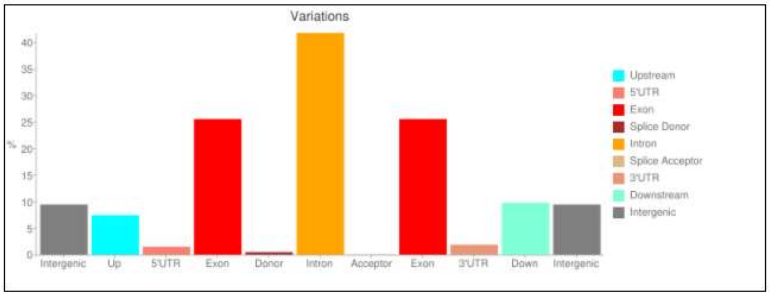

Figura 11: análisis de efectos por tipo y por region

Cómo leer la siguiente tabla:

- Las filas son aminoácidos de referencia y las columnas son aminoácidos cambiados. P.ej. La columna 'E' de la fila 'D' indica cuántos aminoácidos 'D' han sido reemplazados por aminoácidos 'E'.
- Los colores de fondo rojos indican que ocurrieron más cambios (mapa de calor).
- Las diagonales se indican con un color de fondo gris.

	*	-	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
*								3			2										5	2
-																			1			
A		2	134					64							43			3	26	10		
C				16				15									14	5			12	
D			49		28	16		36						3								
E			5		16	1		104			12					3				4		
F				4			20					20						27		18		1
G			20	2	2			151									13	1		4		
H					1				22						69		12					7
I										16		10	1					6	36	17		
K	1	4				26					15			14		6	9		4			
L	4	1					6			10		93			80		54	6		50	1	
M										12		1					10		1	8		
N					14				14		9			28				10	58			
P			7									11			113		5	2	1			
Q						12			36		14				24	13	34					
R	3	3		14				47	15		8	7			23	13	51	14			7	
S			22	7			5	29				6		6	32		35	48	22		1	1
T			22							8	2		5	4	19			18	63			
V			35			5		191		17		17	29							38		
W				4				28				6					14	1				
Y	10			2	8		1		19									104				27

Variants by chromosome

Figura 12: variación por cromosoma

Según el mapeo en UCSC la ubicación es la siguiente:

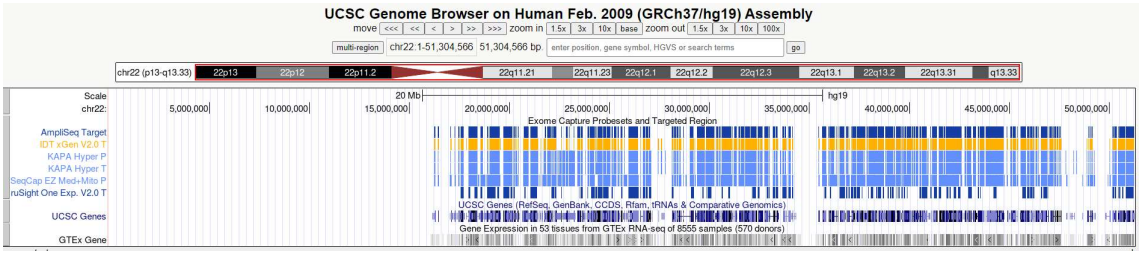

Figura 13: mapeo en el UCSC

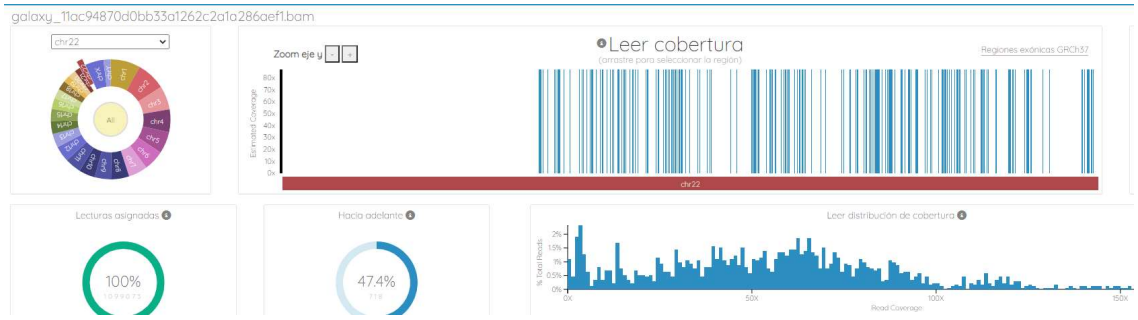

Figura 14: display en bam.iobio

Cuando se analiza las variantes por cromosoma, se identifica que en donde hay mayor cantidad/frecuencia es en el cromosoma 22.

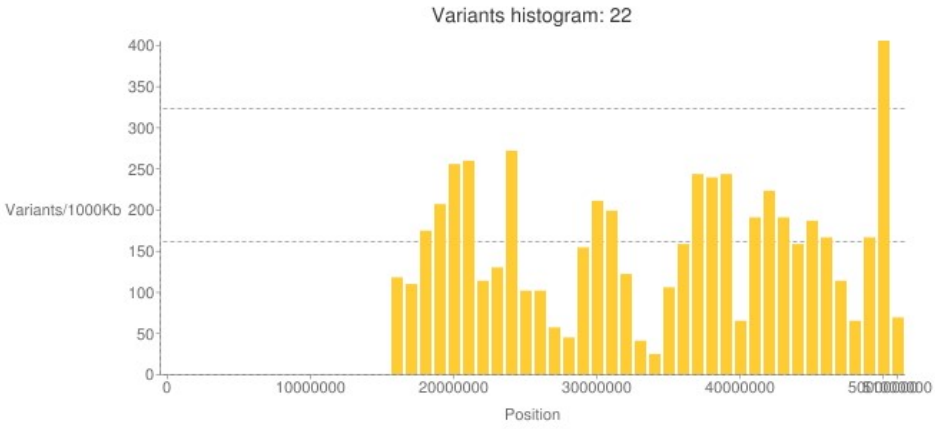

Figura 15: frecuenta por variante por cromosoma

5. Discusión

La secuenciación masiva paralela, conocida como secuenciación de segunda generación o NGS incluye a un conjunto de técnicas con un concepto similar. La capacidad y rapidez de los secuenciadores y el desarrollo continuo de estrategias de testeo más eficaces son parte de la revolución que atravesó la genómica en los últimos años.

Como herramienta, NGS brinda la capacidad de secuenciar a gran escala, con gran versatilidad. Se pueden abordar desde genomas completos hasta paneles reducidos de genes, estudio de fusiones y perfiles de expresión génica por ARN, etc. Las soluciones bioinformáticas permiten analizar los datos genómicos de forma eficaz. Algo que cabe destacar es que NGS, como cualquier otra técnica, tiene sus ventajas y limitaciones, y no necesariamente sustituye a técnicas tradicionalmente establecidas como patrón oro (*gold standard*) en el estudio de ciertas patologías.

Por el contrario, la información genómica que se aporta desde la secuenciación masiva paralela es complementaria a la de otras técnicas, sobre todo en patologías complejas. A lo largo de los años hemos visto cómo el crecimiento y desarrollo de las técnicas de NGS estuvieron asociadas con generación exponencial de conocimiento en el campo de la genómica, lo que en ciertos casos fue llevado al ámbito asistencial por su valor en la clínica.

En los últimos años, con el auge de la medicina de precisión, donde el estudio de un número cada vez mayor de biomarcadores genómicos es un estándar de cuidado, la utilización de este tipo de diagnóstico es de gran utilidad. El costo decreciente para la implementación de estas tecnologías hace pensar que en algunos años este tipo de testeos se realizará de forma rutinaria.

La utilización de herramientas como NGS involucra directa o indirectamente a pacientes, médicos, investigadores, gobiernos, instituciones y seguros de salud, industria farmacéutica y biotecnológica, etc.

Es importante garantizar la calidad de los procesos preanalíticos, analíticos y postanalíticos ligados a NGS, sobre todo en ámbitos asistenciales, por lo tanto, se requiere estandarización y regulación en su utilización, fomentando el trabajo de grupos interdisciplinarios y la formación continua de recursos humanos.

6. Bibliografía

Berkman, P.J., Lai, K., Lorenc, M.T. and Edwards, D. (2012), Next-generation sequencing applications for wheat crop improvement. *American Journal of Botany*, 99: 365-371. https://doi.org/10.3732/ajb.1100309

Hernández, M.: et al (2020). Bioinformatics of next generation sequencing in clinical microbiology diagnosis. *Revista Argentina de Microbiología,* Volume 52, Issue 2, April–June 2020, Pages 150-161.

Lloret-Villas, A., Bhati, M., Kadri, N. K., Fries, R., & Pausch, H. (2021). Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC genomics*, *22*(1), 363. https://doi.org/10.1186/s12864-021-07554-w

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome research*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Nielsen, R., Paul, J., Albrechtsen, A. *et al.* Llamada de genotipos y SNP a partir de datos de secuenciación de próxima generación. *Nat Rev Genet* **12,** 443–451 (2011). https://doi.org/10.1038/nrg2986

Los datos de secuenciación se cargaron en la plataforma web Galaxy y *usamos el servidor público en usegalaxy.org para analizar los datos (Afgan et al. 2016)*