CIB2020

CIB 2020

VIII CONGRESO DE INVESTIGACIÓN BIOMEDICINA

Comparativa de los *Variant Callers*

para la detección de mutaciones somáticas

en análisis de muestra única

i2sysbio INSTITUTE FOR INTEGRATIVE SYSTEMS BIOLOGY

Monfort-Lanzas P[1,2], Arnau V[1], Diaz-Villanueva W[1], Hernando B[2], Martínez-Cadenas C[2]

1. Institute of Integrative Systems Biology (I2Sysbio), University of València and Consejo Superior de Investigaciones Científicas (CSIC), 46980 València, Spain.
2. Genética del cáncer de piel y de la pigmentación humana, Unitat Predepartamental de Medicina, Universitat Jaume I

UNIVERSITAT JAUME I

Introducción

La acumulación de **mutaciones somáticas** en el tejido normal es clave para entender el desarrollo del cáncer. Así pues, su estudio es clave para entender su **transformación** en tejido tumoral (1). Sin embargo el estudio de estas mutaciones es complicado por la composición en mosaico en los tejidos sanos y el ruido de las lecturas obtenidas por NGS.

Los algoritmos para la detección de mutaciones somáticas emplean muestras germinales del paciente para distinguir entre las mutaciones germinales y somáticas pero en muchas ocasiones **no se dispone de muestra de línea germinal** del paciente. Por ello se han desarrollado un gran número de *Variant Callers* (VC) que presentan la opción de **muestra única** (2). Entre los más empleados se encuentran Mutec2, Octopus y Pisces. Si bien, la eficiencia de estos programas para muestra única no está bien caracterizada.

El objetivo de este trabajo es caracterizar la sensibilidad, precisión y FDR de los principales *Variant Callers* de muestra única mediante una base de datos tumoral simulada a partir del exoma de referencia NA12878.

Metodología

Se ha empleado el genoma público **NA12878** (3) que ha sido ampliamente estudiado. Este genoma ya caracterizado permite **simular** una serie de **muestras tumorales** para la comparación de los diferentes programas. Mediante BAMsurgeon se han introducido mutaciones somáticas en posiciones y frecuencias conocidas (4).

Para ello se realizaron los siguientes pasos:

1) Descarga del exoma NA12878 en formato fastq.
2) Alineamiento mediante BWA mem de los fastq frente al genoma de referencia GRCh37 (hg19).
3) Pretratamiento del fichero BAM(Eliminación de duplicados, y realineamiento)
4) Extracción de las regiones homocigotas, con una profundidad de secuenciación mínima de x20, utilizadas como dianas para BAMSurgeon.
5) Obtención de dos muestras tumorales: A. Presenta variantes con una fracción alélica (VAF) de 0.2, 0.1 y 0.05. B. Presenta mutaciones con una VAF de 0.2, 0.1, 0.05, 0.01 y 0.001.
6) Análisis de las muestras mediante Mutect2, Octopus y Pisces.

Resultados

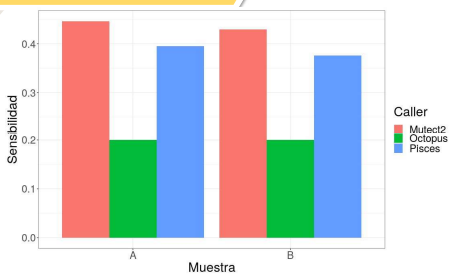

Figura 1. Representación de la sensibilidad de los tres *Variant Callers* en función de la muestra. Se puede observar que la mayor sensibilidad se obtiene empleando Mutect2, y que esta se mantiene con muestra de menor VAF (B).

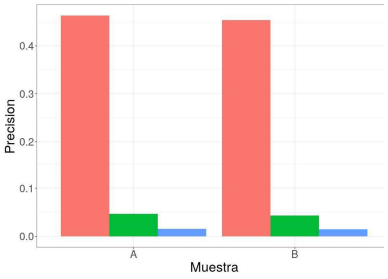

Figura 2. Representación de la precisión de los tres *Variant Callers* en función de la muestra. Se puede observar que la mayor sensibilidad se obtiene empleando Mutect2, se mantiene con muestra de menor VAF (B). También se observa una disminución drástica de la precisión tanto en Octopus como en Pisces, debido a un alto número de Falsos Positivos.

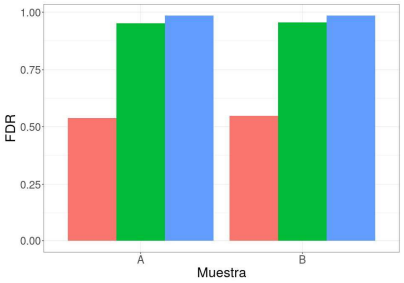

Figura 3. Representación del *False Discovery Rate* (FDR) en los tres *Variant Callers* en función de la muestra. Se puede observar que el menor FDR se obtiene empleando Mutect2. Tanto Octopus como pisces muestran valores muy elevados de FDR debdido al gran número de Falsos Positvos que calsifican.

Conclusión

Para todos los *Variant Callers* comparados, los valores de precisión y sensibilidad obtenidos son inferiores a los que se obtendrían mediante un análisis de muestras pareadas (~95 %). Sin embargo, se observa una variabilidad en función del VC empleado. En este estudio se ha demostrado que los **mejores** resultados se obtienen con **Mutect2**, debido a que el valor de sensibilidad (valor que hace referencia al porcentaje de Verdaderos Positivos) es significativamente mayor .

Aun así, queda patente la necesidad de emplear filtros posteriores para **disminuir el número de falsos positivos** que ha sido incapaz de clasificar correctamente el algoritmo, aproximadamente el 60 %.

Somatic Caller	Octopus		Mutect2		Pisces	
	A	B	A	B	A	B
Verdadero Positivo	330	330	736	708	652	620
Falso Positivo	6700	7265	853	885	41796	41890
Falso Negativo	1320	1320	914	942	998	1030
Sensibilidad	0.2	0.2	0.44	0.42	0.4	0.38
Precisión	0.05	0.04	0.46	0.44	0.02	0.01
FDR	0.95	0.96	0.53	0.55	0.98	0.98

Tabla 1. Resultados obtenidos para Octopus, Mutect2 y pisces, y sus respectivos valores de sensibilidad, precisión y FDR

Referencias

1. Xu, C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Computational and Structural Biotechnology Journal, 16, 15–24
2. Martincorena, I. (2015, May 22). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin.
3. Zook, J. M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology, 32(3), 246–251
4. Meng, J. & Chen, Y.-P. P. (2018). A database of simulated tumor genomes towards accurate detection of somatic small variants in cancer. Plos One, 13(8)