Título: Evaluación comparativa de herramientas bioinformáticas usadas en el mapeo y llamado de variantes para el diagnóstico genético a partir de secuenciación de exoma completo.

Autores:
Diego Saldaña Peñaloza[1,2], Jorge Diaz-Riaño[2], Yenny Gómez[2], Daniel Mahecha[2]

[1] Universidad Nacional de Colombia
[2]Biotecnología y Genética S.A.S - Biotecgen S.A.S

E-mail de contacto: dsaldana@unal.edu.co

Los estudios que comparan el desempeño de herramientas de mapeo y llamado de variantes en el contexto clínico son limitados. Adicionalmente, las pruebas de rendimiento continúan siendo un desafío debido a la escasez de estándares y la falta de definición de consenso para las métricas de rendimiento generando diferencias dramáticas que impactan la toma de decisiones. El objetivo de este trabajo fue comparar el tiempo de procesamiento y el rendimiento de diferentes herramientas usadas en el proceso de mapeo y llamado de variantes a partir de secuenciación de exoma completo. Se procesaron lecturas crudas de secuenciación de exoma completo de la muestra NA12878 con dos herramientas de mapeo: BWA (algoritmo MEM) y Bowtie2, tomando como genoma humano de referencia la versión hg19 del Broad Institute. Los archivos BAM resultantes se procesaron con cinco métodos de llamado de variantes: GATK con red convolucional (CNN) 1D, GATK con CNN 2D, NGSEP, Freebayes y DeepVariant. Se calculó el tiempo total de corrido, así como la Sensibilidad y la tasa de Falsos Positivos Por Millón (FPPM), usando como Gold Standard el VCF de Platinum Genomes para la muestra NA12878. Todos los procesos se ejecutaron en un Servidor Tipo RACK HP DL380P G8, con 2 Procesadores Xeon (12-Core) E5-2697 de 2.7 GHz y Memoria Ram 256 GB, utilizando los parámetros por defecto de cada herramienta (sin paralelización). Para el tiempo de ejecución del mapeo se evidenció una diferencia de 283 minutos entre las herramientas, con 273 minutos para BWA y 556 minutos para Bowtie2. Respecto al llamado de variantes con el BAM resultante de BWA, el mejor tiempo lo obtuvo Freebayes (78.7 minutos), mientras que con el BAM de Bowtie2 lo obtuvo NGSEP (92.7 minutos). En cuanto a las métricas de rendimiento, la mayor sensibilidad en la detección de SNPs heterocigotos se obtuvo con BWA+GATK (CNN 1D), y la mayor sensibilidad en la detección de SNPs homocigotos la obtuvo Bowtie2+Freebayes. La mayor sensibilidad en la detección de Indels heterocigotos la obtuvo BWA+GATK (CNN 1D), mientras que la mayor sensibilidad en la detección de Indels homocigotos corresponde a Bowtie2+DeepVariant. La mejor sensibilidad promedio para SNPs se obtuvo con BWA+Freebayes, seguido de BWA+GATK y BWA+NGSEP. A su vez, la mejor sensibilidad promedio para Indels se obtuvo en BWA+GATK, seguido de Bowtie2+DeepVariant y BWA+Freebayes. La menor FPPM promedio para SNPs se obtuvo con Bowtie2+NGSEP, y para Indels, con Bowtie2+DeepVariant. En conclusión, se encontraron diferencias importantes en los tiempos de ejecución

en las herramientas evaluadas (ignorando la posibilidad de paralelización). Así mismo, al momento de comparar el rendimiento del llamado de variantes hay una mayor diferencia en la sensibilidad para Indels que para SNPs, en consonancia con lo reportado previamente en la literatura. Estos resultados pueden impactar no solamente en el rendimiento diagnóstico sino en el tiempo necesario para lograr un resultado óptimo en el contexto de diagnóstico genético clínico.