Rendimiento en tiempo real para secuenciación Oxford de SARS-CoV-2 en el Valle del Cauca, Colombia

David Esteban Valencia Valencia[1], Nelson Rivera Franco[1,2], Andres Castillo[1], Beatriz Parra Patiño[2], Diana López-Alvarez[1,2,3]

(1) Laboratorio de Técnicas y Análisis Ómicos - TAOLab/CiBioFi, Facultad de Ciencias Naturales y Exactas, Universidad del Valle, Calle 13 No 100-00, Cali, Colombia

(2) Grupo VIREM - Virus Emergentes y Enfermedad, Escuela de Ciencias Básicas, Facultad de Salud, Universidad del Valle, Calle 4B # 36-00, Cali, Colombia

(3) Departamento de Ciencias Biológicas, Facultad de Ciencias Agropecuarias, Universidad Nacional de Colombia, Palmira, Colombia

(diana.lopez.alvarez@correounivalle.edu.co)

La vigilancia genómica del SARS-CoV-2 se ha incorporado como una estrategia efectiva para dar rápida respuesta a los sistemas de salud pública con el fin de llevar a cabo una identificación temprana de variantes emergentes del virus. En este sentido, los protocolos de secuenciación por amplicones usando tecnologías portátiles de Oxford Nanopore Technologies (ONT) han permitido la acumulación a la fecha de más de 6.8 millones de genomas de SARS-CoV-2 según la base de referencia GISAID. Sin embargo, estos protocolos no brindan información del rendimiento esperado de variables como la cobertura mapeada o profundidad mínima en pasos previos al análisis bioinformático. Lo anterior es importante para delimitar los tiempos de secuenciación de muestras y preferiblemente reducir los costos de perdida de reactivos o celdas (flowcell) en entornos con recursos limitados como Latinoamérica. Con el objetivo de estandarizar y predecir la profundidad de las muestras SARS-CoV-2, se validó la dependencia entre los parámetros de ciclo umbral (Ct) de la PCR para el gen N2 (Protocolo CDC), dilución de la muestra, número de lecturas, profundidad y cobertura de secuencia, obtenidos de más de 800 muestras de SARS-CoV-2 para Colombia, principalmente del Valle del Cauca. La secuenciación se basó en el protocolo ARTIC network v3.0 (https://artic.network/ncov-2019) con los kits de secuenciación SQK-LSK109 y de barcodes EXP-NBD196. Se seleccionaron las variables de interés con mayor peso en el análisis para establecer un modelo de Machine Learning a partir del algoritmo con mejor ajuste Gradient Boosting. Se obtuvo que el modelo logro explicar más del 95% de la varianza de la profundidad de secuencia, presentando un 11% de error promedio en las predicciones. En orden de obtener una asignación de linaje o clado en las bases de datos de Pangolin, Nextclade y GISAID, se encontró que se necesitan al menos 18.000 lecturas mapeadas en una muestra determinada para obtener profundidades medias superiores a 200X. Asi, se logró estimar que por cada 50 lecturas totales adicionados en una secuenciación en tiempo real, la profundidad esperada para un barcode aumenta en 1X, dato útil para establecer el tiempo final de corrida de un estudio de secuenciación ONT.