Evaluación del ciclo umbral (Ct) de la qPCR como valor predictivo de la calidad de secuenciación

Consorcio SeqCOVID-SPAIN

Con el fin de cumplir los ambiciosos objetivos propuestos en el proyecto SeqCOVID, hemos emprendido la tarea de secuenciar muestras de pacientes diagnosticados de COVID-19 procedentes de hospitales de referencia de toda la geografía nacional. Gracias a la colaboración de más de 30 hospitales que nos envían extracciones de material genético (ARN), ya hemos podido llevar a cabo la secuenciación de 1440 muestras, una cifra que aumenta semana tras semanas para cubrir todas las fases de la epidemia.

Para la identificación del SARS-CoV-2 a nivel hospitalario se realiza la prueba qPCR, utilizando como material de partida el ARN extraído de la muestra del exudado nasofaríngeo. Uno de los parámetros clave de dicha técnica y que permite evaluar si una muestra es positiva para SARS-CoV-2 es el ciclo umbral o Ct (abreviatura del término en inglés *cycle threshold*), que indica el ciclo a partir del cual se comienza a detectar fluorescencia por encima del ruido de fondo. Por tanto, para un mismo marcador, ciclos de Ct menores se asocian a mayores cantidades de material genético.

El valor de Ct depende de la cantidad de material genético contenido en la muestra, pero existen otros factores que pueden influir en las mediciones de fluorescencia, como la presencia de inhibidores de PCR, o los kits e instrumentos utilizados. En nuestro caso nos interesa la relación inversa de Ct respecto a la cuantía de material genético, puesto que nos puede dar una idea aproximada del éxito de la secuenciación; de tal modo que frente a menores valores de Ct, la cantidad de ARN de partida será mayor, y por tanto podemos esperar una secuenciación más exitosa.

En el contexto en que nos encontramos, definimos el "éxito" de secuenciación como la capacidad de secuenciar una muestra de manera que al menos el 90% del genoma esté cubierto (con una profundidad de al menos 30X). Análisis preliminares nos indican que podemos emplear el parámetro Ct como orientación para seleccionar qué muestras priorizar, y en caso extremo, si proceder o no a su secuenciación.

Relación entre ciclo umbral y cobertura genómica.

Actualmente existe una amplia variedad de kits y diseños para la técnica de qPCR, aún así todos ellos se basan en la amplificación de tres genes: E, N y/o RdRP. En el momento en que se realiza este informe se han secuenciado 1440 muestras, de las cuales 632 (un 44%) presentan el dato de Ct para al menos uno de dichos genes:

- De 492 muestras conocemos el Ct para el gen E
- De 323 muestras conocemos el Ct para el gen N
- Y de 237 muestras conocemos el Ct para el gen RdRP

Hemos escogido el valor de **Ct del gen E** como criterio para la toma de decisiones respecto a la secuenciación. Esto ha sido motivado por dos razones:(i) es el dato que reportan los hospitales con mayor frecuencia, y (ii) de media es el gen que presenta unos valores de Ct

más bajos (24.8 frente a 27.5 para el gen N y 28.3 para el gen RdRP). Este resultado puede ser indicativo de que con el ensayo diseñado para la amplificación del gen E se obtenga una mayor sensibilidad.

Una vez comparados los datos de Ct del gen E con los resultados obtenidos en la secuenciación (en concreto la cobertura genómica conseguida para cada muestra), llegamos a la conclusión de que una alta proporción de muestras con valores de Ct superiores a 33 presentaban un porcentaje de cobertura genómica inferior a 90, por lo que no se consideran aptas para secuenciación (figura 1A). Concretamente, de un total de 492 muestras con Ct conocido para el gen E, 459 tenían un valor de Ct <= 33. Dentro de este rango, 442 muestras (un 96%) alcanzan más del 90% coverage. Sin embargo sólo 14 de 33 muestras (un 42%) con Ct > 33 alcanzan más del 90% coverage.

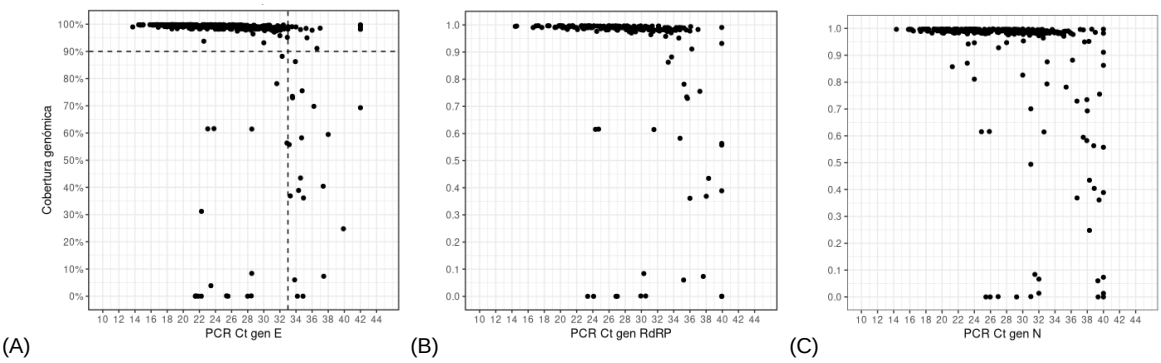


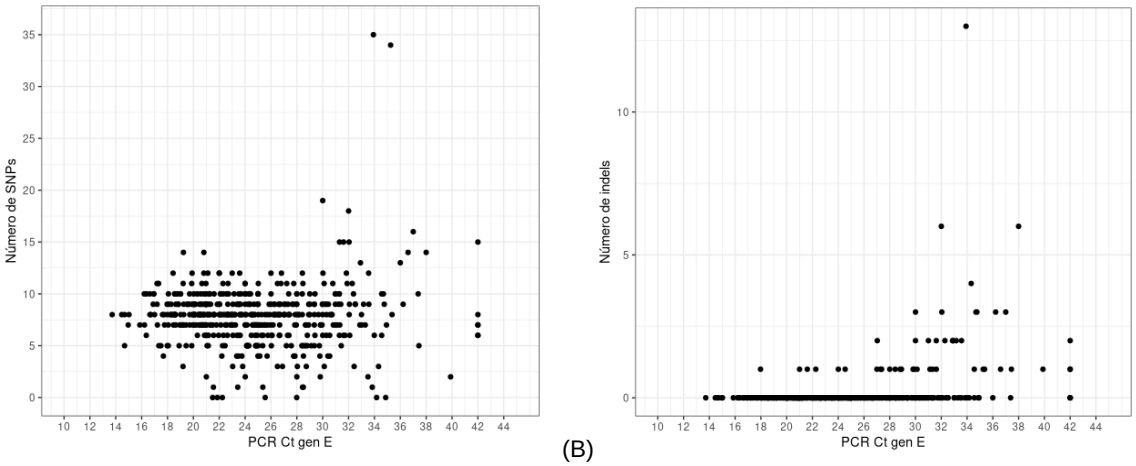
Figura 2. Relación entre Ct (qPCR) y cobertura genómica. (A) Muestra la cobertura genómica que alcanzan las muestras con Ct conocido para el gen E, las líneas discontinuas delimitan los valores de corte establecidos (más del 90% de cobertura como índice de buena calidad de secuenciación y Ct menor que 33). Se observa como una alta proporción de las muestras que supera el límite de Ct < 33 no alcanza el 90% de cobertura genómica. (B)(C) Muestra la cobertura genómica en relación a los valores de Ct de los genes RdRP y N respectivamente.

Por ello se estableció el **punto de corte en Ct gen E <= 33**, tal como señalan las líneas discontinuas de la figura 1A. Esta referencia nos ayuda a decidir qué muestras priorizar para la secuenciación. Cabe puntualizar que todas las muestras, independientemente del valor de Ct gen E, están siendo secuenciadas, dado que en algunas sí se logran una calidad adecuada. A día de hoy estamos empleando este criterio únicamente para dar preferencia a la secuenciación de las muestras que lo cumplan.

Relación entre ciclo umbral y detección de mutaciones

Paralelamente a la influencia que pueda tener el valor de Ct en el coverage, también hemos estudiado otro factor que puede verse afectado durante la secuenciación. Se ha observado que al secuenciar **muestras con Ct altos**, es decir, donde la cantidad de material genético de partida es pequeña, pueden **introducirse artefactos**.

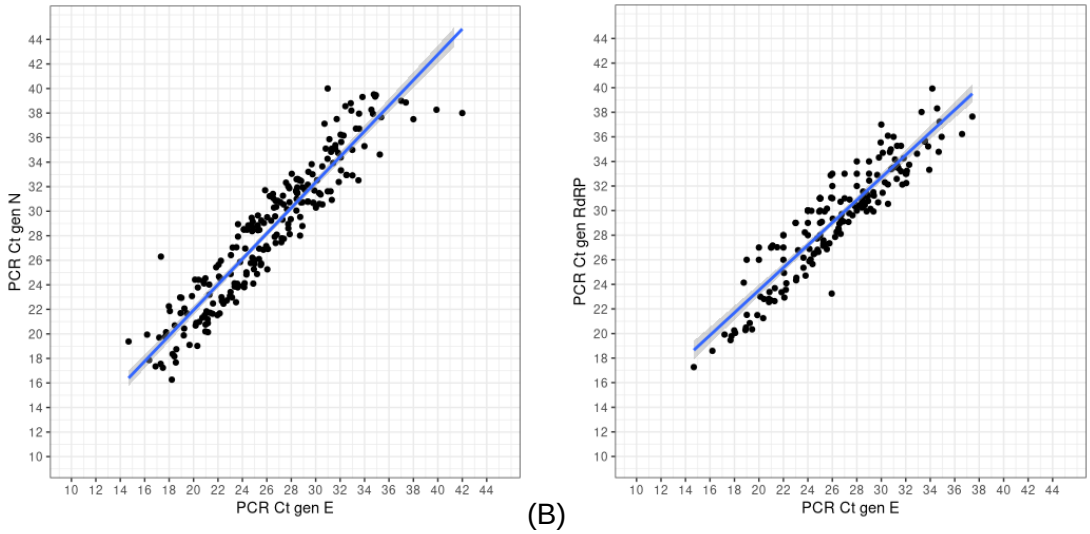
Hemos analizado esta posibilidad en dos puntos clave: el llamamiento de SNPs y la detección de indels. Para ello hemos comparado los valores de Ct gen E con el número de SNPs y el número de indels respectivamente. Tal como se observa en el gráfico de dispersión (figura 2A) esto no parece un problema a la hora de llamar SNPs, puesto que no se observa una correlación clara entre el valor de Ct y la cantidad de SNPs de una muestra (Pearson = 0.06, no significativo con p-val = 0.18). Sin embargo, **sí parece que afecta a los indels** (figura 2B)**,** puesto que, aunque no siempre, y con una correlación baja, las muestras con mayores valores de Ct tienden a presentar más indels en el genoma consenso (Pearson= 0.34, **significativo** p-val < 0.01).



(A) (B)

Figura 2. Relación entre Ct gen E (qPCR) y número de mutaciones detectadas por secuenciación. (A) Representa el número de SNPs encontrados en cada muestra en relación su valor de Ct, se observa una falta de correlación lineal entre ambas variables. (B) En este caso se representa el Ct del gen E frente a la cantidad de indels detectados en cada muestra. Puede verse una tendencia creciente en el número de indels conforme incrementa el valor de Ct de la muestra.

En el caso de no disponer del dato para el gen E, podemos llevar a cabo una interpolación a partir de los valores de Ct para los genes N y RdRP, puesto que existe una correlación lineal entre ellos. El cálculo del coeficiente de correlación de Pearson arroja los siguientes resultados: 0.94 para los Cts de los genes E y N (figura 3A), y 0.91 para los Cts de los genes E y RdRP (figura 3B); en ambos casos con p-val < 0,01. Por tanto podemos concluir que existe una relación lineal directa entre las variables en ambos casos. A continuación se muestran los diagramas de dispersión en los que se puede observar de forma visual dicha correlación.



(A) (B)

Figura 3. Diagramas de dispersión en los que se comparan los valores Ct del gen E con el Ct del gen N (A) y con el Ct del gen RdRP (B) respectivamente. Puede observarse la correlación lineal directa que existe en ambos casos, lo que permite una interpolación de los valores de Ct en caso de carecer del dato de Ct para el gen E.

A medida que se obtengan nuevos datos tendremos la capacidad de llevar a cabo un análisis más robusto, que nos permita comprobar cuán certeros son nuestros resultados preliminares, y así poder establecer con un mayor grado de confianza el rango de Ct que nos asegure una cobertura genómica óptima. Ante la alta incidencia de COVID-19 es importante fijar criterios de estas características que nos ayuden a establecer un orden de prioridad para la secuenciación de muestras. Gracias a ello podemos dinamizar nuestro flujo de trabajo y obtener con mayor celeridad resultados que puedan traducirse en acciones concretas para el control de la pandemia.

Elaborado por los miembros del consorcio: Ana María García Marín, Galo Adrián Goig Serrano. Instituto de Biomedicina de Valencia. Consejo Superior de Investigaciones Científicas.

SeqCOVID-SPAIN es un esfuerzo nacional para secuenciar SARS-CoV-2 en miles de pacientes para entender aspectos epidemiológicos y clínicos de la enfermedad. Más detalles en: seqcovid.csic.es Contacto: icomas@ibv.csic.es

Valencia, 30/06/2020