**Informe Genómica Computacional**

**Desafío**

Sebastian Wolter Salas

**Introducción**

SARS-CoV-2 (o CoV-2)es un virus patogénico altamente transmisible y de carácter zoonótico, registrándose por primera vez en diciembre del 2019 (Ren *et al.,* 2020). Es capaz de infectar a seres humanos, produciendo una gran variedad de síntomas, aunque se sabe que aproximadamente un 20% es capaz de desarrollar síntomas severos, incluyendo neumonía, falla respiratoria e incluso la muerte (Zhang *et al.,* 2020). Este virus se caracteriza por poseer RNA de hebra simple, capaz de reconocer los receptores de la Enzima Convertidora de Angiotensina 2 (ACE2) (Murillo *et al.,* 2020). Aunque en un principio se estimó que esta solo era capaz de afectar los pulmones, más estudios sugieren que existen diversas células del cuerpo humano capaz de interactuar con el virus aunque no necesariamente se presente un sintomatología evidente (Zhang *et al.,* 2020). Debido al alto porcentaje de la población que posee sintomatología leve o no posee sintomatología (asintomáticos) es que existe una gran importancia en términos epidemiológicos para lograr rastrear la presencia de este virus en el ambiente. Además, este es capaz de permanecer infectivo durante varios días en diferentes materiales (Fiorillo *et al.,* 2020). Es por esto, que se ha realizado el metatranscriptoma de diversas fuentes y locaciones de la ciudad de Santiago, Chile, con tal de responder las siguientes preguntas:

¿Podemos detectar CoV-2 en las muestras?

¿Podemos reconstruir los genomas?

¿Podemos detectar a qué linaje pertenecen?

¿Dónde abundan qué variantes de CoV-2? (En qué superficies y dónde en la ciudad)

¿Hay alguna correlación geográfica con respecto al CoV-2 detectado?

Para realizar este análisis y lograr responder estas preguntas, es que se realizó la siguiente metodología.

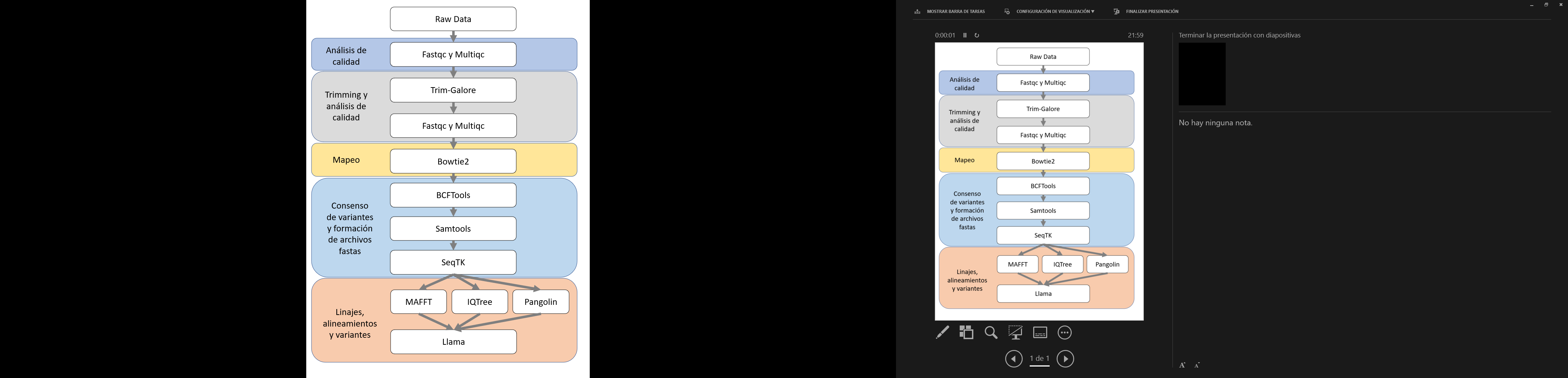
**Metodología**

Para comenzar, se obtuvo la metadata relacionada a los muestreos de los metatranscriptomas utilizados para realizar el análisis de este trabajo, detallado en Tabla 1. Se analizo además el estado y calidad de los metatranscriptomas mediante fastqc (Figuras Suplementarias 1). Las versiones de cada programa esta detallado en Material Suplementario (*Script* documentado). Posteriormente se realizó el *trimming* mediante Trim-Galore!, eliminando todas aquellas secuencias bajas a un Phred promedio menor a 25, de un largo menor a 50 bases, eliminando las bases no reconocidas (bases de carácter “N”). Nuevamente, se realizó el análisis de los archivos de metatranscriptomas mediante fastqc, previo y posterior al “trimming”. Luego de este filtrado de datos, es que se logra observar una leve mejoría en la calidad de los datos, los cuales se encontraban en una buena calidad. Esto es muy probablemente debido a la remoción de los adaptadores de cada *read* obtenida en las muestras. En promedio, se logró remover un ≈ 5% de las *reads* en base a estos criterios (Figuras Suplementarias 2).

**Tabla 1.** Metadata de los metatranscriptomas utilizados en este análisis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Muestra | Latitud | Longitud | Locación | Tipo de lugar | Lugar de muestra | Material del lugar |
| BC-0345795864 | -70,5964 | -33,5200 | Metro Vicuña Mackenna (Comuna La Florida) | Subway station | Escalator handrail | Rubber |
| BC-0345795896 | -70,6053 | -33,4542 | PD335-Parada 1 / (M) Ñuñoa (Comuna Ñuñoa) | Bus stop | Seat | Metal |
| BC-0345797312 | -70,5749 | -33,6096 | Farmacia solidaria Plaza de Puente Alto (Comuna Puente Alto) | Drugstore entrance | Handrail | Metal |
| BC-0345797317 | -70,5999 | -33,4546 | Walmart (Comuna Ñuñoa) | Supermarket Entrance | Handrail | Metal |
| BC-0345797602 | -70,6003 | -33,4538 | Walmart (Comuna Ñuñoa) | Supermarket Entrance | Handrail | Metal |
| BC-0345797632 | -70,5760 | -33,6096 | PF176-Parada 3 / (M) Plaza De Puente Alto (Comuna Puente Alto) | Bus stop | Horizontal pole | Metal |
| BC-0345797636 | -70,6053 | -33,4543 | PD335-Parada 1 / (M) Ñuñoa (Comuna Ñuñoa) | Bus stop | Horizontal pole | Metal |
| BC-0345797641 | -70,6005 | -33,4537 | Walmart (Comuna Ñuñoa) | Supermarket Entrance | Handrail | Painted metal |
| BC-0345797645 | -70,611 | -33,4556 | Home | Home | Control |  |
| BC-0345797686 | -70,6053 | -33,4542 | PD335-Parada 1 / (M) Ñuñoa (Comuna Ñuñoa) | Bus stop | Seat | Metal |
| BC-0345797802 | -70,6052 | -33,4540 | Santander Bank (Comuna Ñuñoa) | Bank | ATM | Metal |
| BC-0345797815 | -70,5994 | -33,5197 | Metro Bellavista de la Florida (Comuna La Florida) | Subway station | ATM | Metal |
| BC-0345797817 | -70,5994 | -33,5197 | Metro Bellavista de la Florida (Comuna La Florida) | Subway station | ATM | Metal |
| BC-0345797839 | -70,5964 | -33,5199 | Metro Vicuña Mackenna (Comuna La Florida) | Subway station | Escalator handrail | Rubber |
| BC-0345798087 | -70,5993 | -33,5197 | Metro Bellavista de la Florida (Comuna La Florida) | Subway station | ATM | Metal |
| BC-0345798943 | -70,597 | -33,5223 | PE1262-Parada 5 / (M) Bellavista De La Florida (Comuna La Florida) | Bus stop | Information board | Metal |
| BC-0345799013 | -70,5745 | -33,6091 | PF512-Parada 2 / (M) Plaza De Puente Alto (Comuna Puente Alto) | Bus stop | Horizontal pole | Metal |
| BC-0361053718 | -70,5989 | -33,4737 | PD64-Parada 1 / Colegio San Marcos (Comuna Macul) | Bus stop | Pole | Metal |
| BC-0361053744 | -70,6071 | -33,4543 | Medical Center (Comuna Ñuñoa) | Practitioner office | Door handle | Metal |
| BC-0361057091 | -70,6071 | -33,4544 | Medical Center (Comuna Ñuñoa) | Practitioner office | Door handle | Metal |

A continuación se realizó el mapeo con Bowtie2 contra la referencia de CoV-2 NC\_045512.2, desde ahora, el genoma de referencia, el cual corresponde a un aislado encontrado en Wuhan en abril del año 2020. Este archivo fue inmediatamente convertido a .bam con su respectivo índice. Luego se analizó mediante bcftools mpileup y bcftools call las posibles variantes de CoV2 encontradas al momento del análisis con la referencia para llegar a una secuencia consenso. Finalmente estos archivos se transforman a un documento tipo fasta. Para lograr analizar los SNPs de cada genoma de CoV2 encontrados en los archivos de metatranscriptoma, era necesario realizar 3 análisis. En primer lugar, se realiza un alineamiento contra la referencia mencionada de cada metatranscriptoma mediante MAFFT. Posteriormente se debe llegar a un consenso de cuales son las cepas encontradas en cada metatranscriptoma utilizando Pangolin. Finalmente se utiliza IQTree para realizar un árbol filogenético. Con eso, se puede realizar un análisis rápido mediante Llama, con tal de determinar cuales son los SNPs de los genomas CoV2 de cada metatranscriptoma. Además, los genomas reconstruidos se analizaron mediante CoVsurver de la plataforma de GISAID para determinar el clado reconocido para cada secuencia. El *script* completo se encuentra en material suplementario 1.



**Figura 1.** Flujo de trabajo utilizado en este estudio.

**Resultados y discusión**

Mediante el análisis de los genomas obtenidos mediante el mapeo en contra de la referencia de los reads de los metatranscriptomas, es que se logro determinar la presencia de CoV-2 en los metatranscriptomas en al menos 14 muestras (Tabla 2). El criterio de exclusión asociado a la detección de genoma de CoV-2 en las muestras fue de 10X de profundidad de secuenciacion posterior a los análisis de mapeo mediante Bowtie2. Debido al tamaño pequeño del genoma de CoV2, es que para asegurar que está presente en la muestra, debe estar secuenciado mediante una gran cantidad de *reads*, por lo que la profundidad se transforma en un criterio relevante para lograr discriminar esta situación. Si bien, sería ideal tener un control negativo para contrastar dichos valores y poder afirmar con mayor certeza la presencia de genoma de CoV-2 en las muestras de transcriptomas, se pueden establecer criterios exigentes con tal de afirmar la posibilidad de encontrarse muestras positivas para este genoma con mayor confianza, razón por el cual se consideró una profundidad de 10X como un umbral sintético adecuado para determinar la detección de la presencia del genoma de este virus. Si bien en estudios como el de Zhang *et al.* (2020) determinan la presencia de CoV-2 sin la necesidad de umbral, ellos extraen el metatranscriptoma directamente desde pacientes positivos para CoV-2, por lo que existe confianza de la presencia de dicha existencia del genoma. En este caso son muestras ambientales, lugar donde se pueden presentar una gran cantidad de diversidad de genomas y transcriptomas provenientes de diversos virus. Es por esto que se prefirió un umbral de detección que va acorde con el porcentaje de cobertura del genoma del virus. Con respecto a la reconstrucción del genoma, se tomaron en cuenta los criterios de exclusión encontrados en la [página de GISAID](https://www.epicov.org/epi3/frontend#43036b). Principalmente, es necesario obtener mas de 29 mil bases mapeadas y además poseer menos de un 5% de “N” en la secuencia resultante para considerar que se encuentra el genoma de CoV-2 reconstruido. Mediante estos criterios, es que se logro reconstruir 14 genomas en total (Tabla 2). De hecho, existe una relación clara en base a aquellas muestras que presentaron mayor profundidad de mapeo del virus con respecto al mayor porcentaje de cobertura y en consecuencia, poseer mayor cantidad de bases mapeadas (Tabla 2).

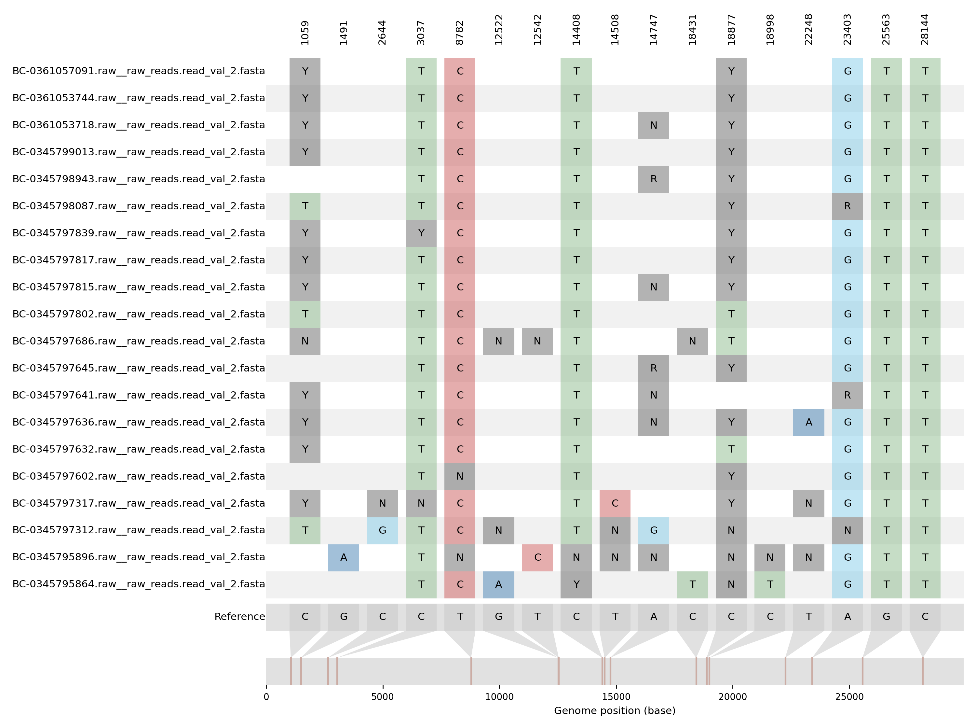
Se analizo el linaje al cual podrían pertenecer los genomas reconstruidos mediante el mapeo contra la referencia mencionada. En base a lo anterior, es que mediante la plataforma de Pangolin, se logró detectar el linaje para cada una de las muestras, aunque con diversas probabilidades de asignación correcta (Tabla 2). En base a los criterios de exclusión de GISAID mencionados previamente, se analizaron los genomas que fueron efectivamente reconstruidos, debido a que la deficiencia en la cobertura podría eventualmente resultar en falsos positivos para la selección del linaje asignado. Mediante Pangolin, es que se logró detectar el linaje B.1 en todas las muestras, aunque con diferencias en la probabilidad de asignación. De hecho, los genomas seleccionados como reconstruidos, se logró determinar una distribución entre 0.36 hasta 0.91, siendo predominante la alta probabilidad de asignación correcta del linaje al genoma reconstruido. En mayor detalle, se encontraron 10 genomas con un 0.91 de probabilidad de asignación, mientras que 0.50 se encontró 1 vez y 0.36 se encontró 3 veces. Con respecto a aquellos que resultaron en una menor probabilidad, puede estar otorgado debido a que la cobertura incompleta de los genomas. Al no ser 100% esta cobertura, pueden existir regiones sin mapear, las cuales son específicas para detectar el linaje al que pertenecen. Al perder estas regiones, la plataforma evita la correcta identificación del linaje del genoma, disminuyendo la probabilidad de una correcta correlación entre estos. Además del análisis de Pangolin, se utilizó la plataforma CoVsurver de GISAID, el cual determino que el clado correspondiente a 16 de los 20 genomas obtenidos serian del clado GH. Con respecto a los 14 genomas reconstruidos, es que 4 de no pudieron ser identificados. Debido a que las plataformas se basan en criterios diferentes, la falta de ciertas regiones en el genoma puede afectar la clasificación de forma distinta, por lo que pueden generar resultados dispares. En base a lo anterior, se logró determinar el linaje de 14 genomas y el clado de 10 en las plataformas mencionadas. Con respecto al linaje B.1 encontrado en los genomas reconstruidos, cabe destacar su procedencia europea, correspondiente a un brote de Italia. Han existido reportes de cepas europeas ingresadas a Chile, por lo que se podría hipotetizar la entrada de este linaje mediante personas que viajaron desde Europa a Chile. Si analizamos desde los clados de la plataforma de GISAID, el clado GH se conformó en febrero de 2020, expandiéndose en gran parte de América del Norte, Europa y América del Sur, incluyendo [Chile](https://www.cov2.cl/#/) (Mercatelli y Giorgi, 2020), por lo que se correlaciona encontrar este clado en particular en las muestras.

**Tabla 2.** Resultados estadísticos y detección de linaje y clado de genomas reconstruidos.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Muestra | Porcentaje de *reads* mapeado (%) | | Profundidad | Porcentaje de genoma cubierto | Bases totales mapeadas | Linaje | Probabilidad de asignación (Pangolin) | Clado GISAID |
| BC-0345795864 | 0.0025 | 6.79 | | 89.85 | 26839 | B.1 | 0.91 | **GH** |
| BC-0345795896 | 0.0009 | 3.35 | | 64.20 | 19171 | B.1 | 0.91 | **GH** |
| BC-0345797312 | 0.0014 | 4.64 | | 81.77 | 24416 | B.1 | 0.36 | **GH** |
| BC-0345797317 | 0.0015 | 5.26 | | 81.69 | 24399 | B.1 | 0.50 | **GH** |
| BC-0345797602 | 0.0027 | 9.06 | | 95.59 | 28548 | B.1 | 0.91 | **GH** |
| BC-0345797632 | **0.0128** | **39.34\*** | | **99.82** | **29818** | **B.1** | **0.50** | **GH** |
| BC-0345797636 | **0.0043** | **13.79\*** | | **99.94** | **29555** | **B.1** | **0.91** | **Otra** |
| BC-0345797641 | **0.0144** | **45.01\*** | | **99.93** | **29852** | **B.1** | **0.91** | **Otra** |
| BC-0345797645 | **1.1467** | **2660.99\*** | | **99.94** | **29853** | **B.1** | **0.36** | **GH** |
| BC-0345797686 | 0.0011 | 4.27 | | 80.25 | 23972 | B.1 | 0.91 | **GH** |
| BC-0345797802 | **0.3427** | **703.90\*** | | **99.96** | **29858** | **B.1** | **0.36** | **GH** |
| BC-0345797815 | **0.0191** | **50.36\*** | | **99.87** | **29833** | **B.1** | **0.36** | **GH** |
| BC-0345797817 | **0.0309** | **90.21\*** | | **99.94** | **29854** | **B.1** | **0.91** | **GH** |
| BC-0345797839 | **0.0132** | **41.20\*** | | **99.93** | **29851** | **B.1** | **0.91** | **GH** |
| BC-0345798087 | **0.0094** | **22.20\*** | | **99.68** | **29775** | **B.1** | **0.91** | **Otra** |
| BC-0345798943 | **0.0056** | **16.71\*** | | **98.31** | **29366** | **B.1** | **0.91** | **Otra** |
| BC-0345799013 | **0.0174** | **53.47\*** | | **99.96** | **29860** | **B.1** | **0.91** | **GH** |
| BC-0361053718 | **0.0180** | **49.55\*** | | **99.82** | **29814** | **B.1** | **0.91** | **GH** |
| BC-0361053744 | **0.0625** | **181.37\*** | | **99.96** | **29859** | **B.1** | **0.91** | **GH** |
| BC-0361057091 | **0.0630** | **179.44\*** | | **99.42** | **29697** | **B.1** | **0.91** | **GH** |

\*Secuencias que lograron superar una profundidad de 10X.

Se realizo el análisis de las SNPs de los genomas encontrados en este trabajo, incluyendo aquellos reconstruidos y no reconstruidos. Mediante llama es que se logró identificar SNPs específicos que fueron mutados con respecto a la referencia (Figura 3). Se logro detectar SNPs presentes en todas los genomas reconstruidos, como las transiciones de “C” a “T” en 3037, 14408 y 28144, la transición de “A” a “G” en 23403 y la transversión de “G” a “T” en 25563. A pesar de las diferencias entre SNPs de algunos genomas, es que Pangolin los agrupo al linaje B.1. Es por esto que es necesario complementar la probabilidad de asignación, en conjunto al porcentaje de cobertura y profundidad, con tal de determinar específicamente una variable.



**Figura 3.** Asignación de SNPs para cada genoma reconstruido.Para cada genoma se señala la base mutada en conjunto a la posición del genoma. Para cada base se tiene asignado un color, siendo para Timina verde, Adenina azul, Guanina celeste y Citocina Rojo. Los nucleótidos con la letra “N” grises son aquellos que no lograron ser identificados, mientras que aquellos con la letra “Y” hace referencia a Citosina o Guanina.

Con respecto a la abundancia de las variantes de CoV-2, es que se logran detectar principalmente en superficies de metal, a excepción de la muestra BC-0345797839, la cual fue obtenida desde caucho (Tabla 1). Esto va acorde a lo mencionado por Fiorillo *et al.* (2020). En este estudio se realizó una revisión de diversos artículos científicos de aquellos sitios que están mayormente expuestos a recibir y mantener los virus en estado infeccioso, concluyendo que los metales y los plásticos son capaces de mantener el virus activo, incluso por 9 días a temperatura ambiente, mientras que el caucho durante 3 días. Es por esto que existe una alta relación a encontrar una mayor cobertura y profundidad en aquellas muestras obtenidas precisamente de superficies de metal.

Además, en este trabajo se logro detectar una mayor cobertura de genomas desde muestras provenientes de ATMs, manijas de puertas, y en menor medida de pasamanos y postes, lugares donde existe un alto contacto manual. De hecho, es importante considerar que estas muestras provienen principalmente de estaciones de metro (BC-0345797815, BC-0345797817, BC-0345797839, BC-0345798087) y paraderos de buses (BC-0345797632, BC-0345797636, BC-0345798943, BC-0345799013, BC-0361053718), por lo que estaría correlacionado con el alto flujo de personas que reciben constantemente estos lugares. También existen dos muestras relacionadas a un centro médico (BC-0361053744, BC-0361057091), donde se detecta con una alta profundidad el genoma de CoV-2, situación que estaría relacionada a la atención de pacientes infectados con este virus.

Finalmente, existen 6 muestras con alta cobertura que provienen de Ñuñoa, comuna en la que se encontraba con una alta cantidad de personas infectadas en el momento de la toma de muestra. Esta comuna entro el jueves 26 de marzo a cuarentena debido a la alta cantidad de personas de esa localidad infectadas con CoV-2. Otra comuna afectada al comienzo de la pandemia en Chile fue La Florida, lugar donde se encontraron 5 muestras positivas, comuna también bastante afectada en abril-mayo de este año. Ya para el 16 de abril se reportaba como una de las comunas con una alta cantidad de personas infectadas, para entrar a cuarentena el 8 de mayo. En conjunto a lo anterior, se encontraron 2 muestras positivas provenientes de Puente Alto, comuna que entro a cuarentena el 9 de abril debido a la alta cantidad de contagios. Dado el momento de la toma de muestras (abril-mayo), la zona poniente estaba siendo afectada gravemente con la cantidad de infectados de CoV-2, debido un deficiente rastreo de personas infectadas, por lo que encontrar genomas completos dentro de estas muestras transcriptomas esta correlacionada con el lugar desde donde se obtienen.

**Conclusión**

Mediante el análisis metatranscriptómico de diferentes superficies de Santiago, se logro detectar y reconstruir el genoma de CoV-2 en 14 muestras de un total de 20. Con esto, además se pudo identificar el linaje y clado de estas mediante las plataformas de Pangolin y GISAID. Con respecto a la relación geográfica, las muestras mas abundantes se encuentran relacionadas con lugares de alto tránsito y en comunas con un alto nivel de contagios, ademas de encontrarse en superficies de alto contacto, como ATMs, manijas de puertas y postes.

**Referencias**

Fiorillo, L., Cervino, G., Matarese, M., D’Amico, C., Surace, G., Paduano, V., Cicciù, M. (2020). COVID-19 Surface Persistence: A Recent Data Summary and Its Importance for Medical and Dental Settings. International Journal of Environmental Research and Public Health, 17(9), 3132.

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Frontiers in Microbiology, 11.

Murillo, J., Villegas, L. M., Ulloa-Murillo, L. M., & Rodríguez, A. R. (2020). Recent Trends on Omics and Bioinformatics Approaches to Study SARS-CoV-2: A Bibliometric Analysis and mini-review. Computers in Biology and Medicine, 104162. Advance online publication.

Ren, L. L., Wang, Y. M., Wu, Z. Q., Xiang, Z. C., Guo, L., Xu, T., Jiang, Y. Z., Xiong, Y., Li, Y. J., Li, X. W., Li, H., Fan, G. H., Gu, X. Y., Xiao, Y., Gao, H., Xu, J. Y., Yang, F., Wang, X. M., Wu, C., Chen, L., Wang, J. W. (2020). Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chinese medical journal, 133(9), 1015–1024.

Zhang, H., Ai, J. W., Yang, W., Zhou, X., He, F., Xie, S., Zeng, W., Li, Y., Yu, Y., Gou, X., Li, Y., Wang, X., Su, H., Xu, T., & Zhang, W. (2020). Metatranscriptomic Characterization of COVID-19 Identified A Host Transcriptional Classifier Associated With Immune Signaling. Clinical infectious diseases: an official publication of the Infectious Diseases Society of America, ciaa663. Advance online publication.

Wu,F., Zhao,S., Yu,B., Chen,Y.M., Wang,W., Song,Z.G., Hu,Y., Tao,Z.W., Tian,J.H., Pei,Y.Y., Yuan,M.L., Zhang,Y.L., Dai,F.H., Liu,Y., Wang,Q.M., Zheng,J.J., Xu,L., Holmes,E.C., Zhang,Y.Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579, (7798), 265-269.